

DETECTION OF CORE-PERIPHERY STRUCTURE IN NETWORKS USING SPECTRAL METHODS AND GEODESIC PATHS

MIHAI CUCURINGU ^{*}, PUCK ROMBACH [†], SANG HOON LEE [‡], AND MASON A. PORTER [§]

December 11, 2015

Abstract. We introduce several novel and computationally efficient methods for detecting “core-periphery structure” in networks. Core-periphery structure is a type of mesoscale structure that includes densely-connected core vertices and sparsely-connected peripheral vertices. Core vertices tend to be well-connected both among themselves and to peripheral vertices, which tend not to be well-connected to other vertices. Our first method, which is based on transportation in networks, aggregates information from many geodesic paths in a network and yields a score for each vertex that reflects the likelihood that vertex is a core vertex. Our second method is based on a low-rank approximation of a network’s adjacency matrix, which can often be expressed as a tensor-product matrix. Our third approach uses the bottom eigenvector of the random-walk Laplacian to infer a coreness score and a classification into core and peripheral vertices. Additionally, we design an objective function to (1) help classify vertices into core or peripheral vertices and (2) provide a goodness-of-fit criterion for classifications into core versus peripheral vertices. To examine the performance of our methods, we apply our algorithms to both synthetically-generated networks and a variety of real-world data sets.

Keywords. Networks, core-periphery structure, shortest-path algorithms, low-rank matrix approximations, graph Laplacians, spectral algorithms.

1. Introduction. Network science has grown explosively during the past two decades [58], and myriad new journal articles on network science appear every year. One focal area in the networks literature is the development and analysis of algorithms for detecting local, mesoscale, and global structures in various types of networks [31, 67]. Mesoscale features are particularly interesting, as they arise neither at the local scale of vertices (i.e., nodes) and edges nor at the global scale of summary statistics. In the present paper, we contribute to research on mesoscale network structures by developing and analyzing new (and computationally-efficient) algorithms for detecting a feature known as *core-periphery structure*, which consists of densely-connected *core* vertices and sparsely-connected *peripheral* vertices.

The importance of the investigation of mesoscale network structures is widely acknowledged [31, 67], but almost all of the research on this topic concerns a specific type of feature known as *community structure*. In studying community structure, one typically employs some algorithm to detect sets of vertices called *communities* that consist of vertices that are densely connected to each other, such that the connection density between vertices from different communities is comparatively sparse [31, 33, 60, 67]. A diverse array of methods exist, and they have been applied to areas, such as committee networks in political science [66], friendship networks [35, 80], protein-protein interaction networks [18, 50], functional brain networks [7], and mobile phone networks [61]. Popular methods include the optimization of a quality function called “modularity” [57, 59, 60], spectral partitioning [38, 78], dynamical approaches based on random walkers or other dynamical systems [3, 41, 64, 65, 72], local methods such as *k*-clique percolation [62], and more. Most community-detection methods require a vertex to belong to a distinct community, but several methods also allow the detection of overlapping communities (see, e.g., [1, 4, 41, 62]).

Core-periphery structure is a mesoscale feature that is rather different from community structure. The main difference is that core vertices are well-connected to peripheral vertices, whereas the standard perspective on community structure views communities as nearly decomposable modules (which leads to trying to find the best block-diagonal fit to a network’s adjacency matrix) [69, 85]. Core-periphery structure and community structure are thus represented by different types of block models [41, 63]. The quantitative investigation of core-periphery structure has a reasonably long history [24], and qualitative notions of core-periphery structure have long been considered in fields such as international relations [17, 76, 79, 83], sociology [27, 48], and economics [47] (and have been examined more recently in applications such as neuroscience [8], transportation [49], and faculty movements in academia [20]), but the study of core-periphery structure remains poorly developed — especially in comparison to the study of community structure [31, 67]. Most investigations of core-periphery structure tend to use the perspective that a

^{*}Department of Mathematics, UCLA, Los Angeles, CA (mihai@math.ucla.edu). This work was initiated while the author was affiliated with the Program in Applied and Computational Mathematics (PACM) at Princeton University, Princeton, NJ.

[†]Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford, UK; and Department of Mathematics, UCLA, Los Angeles, CA.

[‡]School of Physics, Korea Institute for Advanced Study, Seoul, Korea; and Integrated Energy Center for Fostering Global Creative Researcher (BK 21 plus) and Department of Energy Science, Sungkyunkwan University, Suwon, Korea; and Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford, UK

[§]Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford, UK; and CABDyN Complexity Centre, University of Oxford, Oxford, UK.

network’s adjacency matrix has an intrinsic block structure (which is different from the block structure from community structure) [12, 23, 69]. Very recently, for example, Ref. [85] identified core–periphery structure by fitting a stochastic block model (SBM) to empirical network data using a maximum likelihood method, and the SBM approach in Ref. [63] can also be used to study core–periphery structure. Importantly, it is possible to think of core–periphery structure using a wealth of different perspectives, such as overlapping communities [84], k -cores [39], network capacity [25], and random walks [70]. The notion of “nestedness” [6] from ecology is also related to core–periphery structure, although establishing an explicit and direct connection between these ideas appears to be an open problem. The main contribution of the present paper is the development of novel algorithms for detecting core–periphery structure. Our aim is to develop algorithms that are both computationally efficient and robust to high levels of noise in data, as such data can lead to a blurry separation between core vertices and peripheral vertices.

The rest of this paper is organized as follows. In Section 2, we give an introduction to the notion of core–periphery structure and briefly survey a few of the existing methods to detect such structure. In Section 3, we introduce the PATH-CORE method, which is based on computing shortest paths between vertices of a network, for detecting core–periphery structure. In Section 4, we introduce an objective function for detecting core–periphery structure that leverages our proposed algorithms and helps in the classification of vertices into a core set and periphery set. In Section 5, we propose the spectral method LOWRANK-CORE, which detects core–periphery structure by considering the adjacency matrix of a network as a low-rank perturbation matrix. In Section 6, we investigate two Laplacian-based methods (LAP-CORE and LAPSGN-CORE) for computing core–periphery structure in a network, and we discuss related work in community detection that uses a similar approach. In Section 7, we compare the results of applying the above algorithms using several synthetically-generated networks and real-world networks. Finally, we summarize and discuss our results in Section 8, and we also discuss several open problems and potential applications. In Appendix 1, we detail the steps of our proposed PATH-CORE algorithm for computing the Path-Core scores, and we include an analysis of its computational complexity. In Appendix 2, we discuss the spectrum of the random-walk Laplacian of a graph and (of the random-walk Laplacian of its complement). In Appendix 3, we detail an experiment with artificially planted high-degree peripheral vertices that illustrates the sensitivity of a degree-based method (which we call DEGREE-CORE and which uses vertex degree as a proxy to measure coreness) to such outlier vertices. Finally, in Appendix 4, we calculate Spearman and Pearson correlation coefficients between the coreness scores that we obtain from the different methods across several real-world networks.

2. CORE-SCORE: Density-Based Core–Periphery Structure in Networks. The best-known quantitative treatment of core–periphery structure was introduced by Borgatti and Everett [12], who developed algorithms for detecting discrete and continuous versions of core–periphery structure in weighted, undirected networks. (For the rest of the present paper, note that we will use the terms “network” and “graph” interchangeably.) Their discrete methods start by comparing a network to an ideal block matrix in which the core is fully connected, the periphery has no internal edges, and the periphery is well-connected to the core.

Borgatti and Everett’s main algorithm for finding a discrete core–periphery structure assigns each vertex either to a single “core” set of vertices or to a single “periphery” set of vertices. One seeks a vector \mathbf{C} of length n whose entries are either 1 or 0, depending on whether or not the associated vertex has been assigned to the core (1) or periphery (0). We let $H_{ij} = 1$ if $C_i = 1$ (i.e., vertex i is assigned to the core) or $C_j = 1$ (i.e., vertex j is assigned to the core), and we otherwise let $H_{ij} = 0$ (because neither i nor j are assigned to the core). We define $\rho_C = \sum_{i,j} A_{ij} H_{ij}$, where A (with elements A_{ij}) is the adjacency matrix of the (possibly weighted) network G . Borgatti and Everett’s algorithm searches for a value of ρ_C that is high compared to the expected value of ρ if \mathbf{C} is shuffled such that the number of 0 and 1 entries is preserved but their order is randomized. The final output of the method is the vector \mathbf{C} that gives the highest z -score for ρ_C . In a variant algorithm for detecting discrete core–periphery structure, Borgatti and Everett still let $H_{ij} = 1$ if both C_i and C_j are equal to 1 and let $H_{ij} = 0$ if neither i nor j are assigned to the core, but they now let $H_{ij} = a \in [0, 1]$ if either $C_i = 1$ or $C_j = 1$ (but not both). To detect a continuous core–periphery structure [12], Borgatti and Everett assigned a vertex i a core value of C_i and let $H_{ij} = C_i C_j$. A recent method that builds on the continuous notion of core–periphery structure from [12] was proposed in [69]. It calculates a CORE-SCORE for weighted, undirected networks; and it has been applied (and compared to community structure) in the investigation of functional brain networks [8].

The method of core–periphery detection in the popular network-analysis software UCINET [13] uses the so-called *minimum residual* (MINRES) method [23], which is a technique for factor analysis. One uses factor analysis to describe observed correlations between variables in terms of a smaller number of unobserved variables called the

“factors” [26]. MINRES aims to find a vector C that minimizes

$$S(A, \mathbf{C}) = \sum_{i=1}^n \sum_{j \neq i} (A_{ij} - C_i C_j)^2,$$

where $C_i \geq 0$ for all vertices i . One ignores the diagonal elements of the network’s adjacency matrix. Additionally, because $\mathbf{C}\mathbf{C}^T$ is symmetric, this method works best for undirected networks G . For directed networks, one can complement the results of MINRES with a method based on a singular value decomposition (SVD) [14]. In practice, UCINET reports $\mathbf{C}/\sqrt{\sum_i C_i^2}$.

In Ref. [84], Yang and Leskovec argued that core–periphery structure can arise as a consequence of community structure with overlapping communities. They presented a so-called *community-affiliation graph model* to capture dense overlaps between communities. In their approach, the likelihood that two vertices are adjacent to each other is proportional to the number of communities in which they have shared membership. Della Rossa et al. recently proposed a method for detecting a continuous core–periphery profile of a (weighted) network by studying the behavior of a random walker on a network [70]. Approaches based on random walks and other Markov processes have often been employed in the investigation of community structure [41, 64, 65, 72], and it seems reasonable to examine them for other mesoscale structures as well. Very recently, Ref. [85] identified core–periphery structure by fitting a stochastic block model to empirical network data using a maximum likelihood method. The review article [24] discusses several other methods to detect core–periphery structure in networks.

3. PATH-CORE: Transport-Based Core–Periphery Detection via Shortest Paths in a Network. In transportation systems, some locations and routes are much more important than others. This motivates the idea of developing notions of core–periphery structure that are based on transportation. In this section, we restrict our attention to undirected and unweighted networks, although we have also examined transport-based core–periphery structure in empirical weighted and directed networks [49].

The first transport-based algorithm that we propose for detecting core–periphery structure is reminiscent of *betweenness centrality* (BC) in networks [2, 32, 56]. One seeks to measure the extent to which a vertex controls information that flows through a network by counting the number of shortest paths (i.e., “geodesic” paths) on which the vertex lies between pairs of other vertices in the network. Geodesic vertex betweenness centrality is defined as

$$(1) \quad B_C(i) = \sum_{j,k \in V(G) \setminus i} \frac{\sigma_{jk}(i)}{\sigma_{jk}},$$

where σ_{jk} is the number of different shortest paths (i.e., the “path count”) from vertex j to vertex k , and $\sigma_{jk}(i)$ is the number of such paths that include vertex i . Our approach also develops a scoring methodology for vertices that is based on computing shortest paths in a network. Such a score reflects the likelihood that a given vertex is part of a network’s core. Instead of considering shortest paths between all pairs of vertices in a network, we consider shortest paths between pairs of vertices that share an edge *when that edge is excluded from the network*. More precisely, we calculate

$$(2) \quad \text{PATH-CORE}(i) = \sum_{(j,k) \in E(V(G) \setminus i)} \frac{\sigma_{jk}(i)|_{G \setminus (j,k)}}{\sigma_{jk}|_{G \setminus (j,k)}},$$

where $\sigma_{jk}(i)|_{G \setminus (j,k)}$ and $\sigma_{jk}|_{G \setminus (j,k)}$ are defined, respectively, as the path counts σ_{jk} and $\sigma_{jk}(i)$ in the graph $G \setminus (j,k)$, and $E(X)$ denotes the edge set induced by the vertex set X . The network $G \setminus (j,k)$ denotes the subgraph of G that one obtains by removing the edge $(j,k) \in E$. Alternatively, one can define the PATH-CORE score of a vertex i as the betweenness centrality of this vertex when considering paths only between pairs of adjacent vertices j and k , but for which the edge e_{jk} incident to the two vertices is discarded.

In passing, we mention the approach by Valente and Fujimoto [82] for deriving measures for “bridging” in networks based on the observation that edges that reduce distances in a network are important structural bridges. In their measure, for which they used a modification of closeness centrality, one systematically deletes edges and measures changes in the resulting mean path lengths. See also the recent paper [29] about bridging centrality.

In Section 3.1, we explain the intuition behind the proposed PATH-CORE algorithm, and we examine its performance on several synthetic networks. In Section 3.2, we comment on a randomized version of the PATH-CORE algorithm that samples a subset of edges in a graph and computes shortest paths only between the endpoints of the associated vertices.

3.1. PATH-CORE. Let $G(V, E)$ be a graph with a vertex set V of size n (i.e., there are $|V| = n$ vertices) and an edge set E of size m . The set of core vertices is V_C (and its size is n_c)¹, and the set of peripheral vertices is V_P (and its size is n_p). Suppose that a network (i.e., a graph) contains exactly one core set and exactly one peripheral set, and that these sets are disjoint: $V_C \cup V_P = V$ and $V_C \cap V_P = \emptyset$. The goal of the PATH-CORE algorithm is to compute a score for each vertex in the graph G that reflects the likelihood that that vertex belongs to the core. In other words, high-scoring vertices have a high probability of being in the core, and low-scoring vertices have a high probability of being in the periphery. Throughout the paper, we use the term “PATH-CORE scores” to indicate the scores that we associate with a network’s vertices by using the PATH-CORE algorithm.

We illustrate our methodology in the context of a generalized block model, such as the one in Table 1, where the submatrices $A_{\{CC\}}$, $A_{\{CP\}}$, and $A_{\{PP\}}$ represent the interactions between a pair of core vertices, a core vertex and a peripheral vertex, and a pair of peripheral vertices, respectively. Suppose that $A_{\{CC\}}$ and $A_{\{PP\}}$ are adjacency matrices that we construct using the Erdős-Rényi random graph models² $G(n_c, p_{cc})$ and $G(n_p, p_{pp})$, respectively, and that $A_{\{CP\}}$ is the adjacency matrix of a random bipartite graph $G(n_c, n_p, p_{cp})$ in which each edge that is incident to both a core and peripheral vertex is present with independent probability p_{cp} . In the context of the above block model, core-periphery structure arises naturally in instances of the above ensemble for which $p_{cc} \geq p_{cp} > p_{pp}$ or $p_{cc} > p_{cp} \geq p_{pp}$. The above family of random networks, which we denote by $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$, was also considered in Ref. [69]. It contains exactly one set of core vertices, and the remaining vertices are peripheral vertices. More complicated core-periphery structures can also occur [69], such as a mix of (possibly hierarchical) community structures and core-periphery structures.

$A_{\{CC\}}$	$A_{\{CP\}}$
$A_{\{CP\}}$	$A_{\{PP\}}$

TABLE 1

Block model for the ensemble of graphs $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$. Note that $p_{cc} \geq p_{cp} > p_{pp}$ or $p_{cc} > p_{cp} \geq p_{pp}$.

We now present the intuition behind the PATH-CORE algorithm and the reason that the resulting PATH-CORE score is a good indicator of the likelihood that a vertex is in the core or in the periphery. If i and j are adjacent core vertices, then it is likely that shortest paths between i and j consist entirely of other core vertices. If $i \in V_C$ and $j \in V_P$, then a shortest path between i and j should also mostly contain core vertices. Finally, even when $i, j \in V_P$, it is still likely that a shortest path between i and j is composed of many core vertices and few peripheral vertices. Intuitively, once a shortest path reaches the set V_C , it is likely to stay within the core set V_C until it returns to the periphery set V_P and reaches the terminal vertex j , because $p_{cc} \geq p_{cp} \geq p_{pp}$. To summarize, we expect core vertices to be on many shortest paths in a graph, whereas peripheral vertices should rarely be on such shortest paths. In other words, because shortest paths between a pair of core vertices are the ones that should on average contain the largest fraction of vertices that are in the core, we find that oversampling such paths is an effective way to extract core parts of a graph. Importantly, it is not sufficient in general to simply use a quantity like weighted BC. For example, for a stock-market correlation network that was examined in Ref. [49], weighted BC cannot distinguish the importance of vertices at all, whereas coreness measures (in particular, CORE-SCORE and PATH-CORE) are able to successfully determine core vertices.

To illustrate the effectiveness of the PATH-CORE algorithm, we consider (see Fig. 1) several instances of the random-graph ensemble $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with $p_{cc} > p_{cp} > p_{pp}$. Let $\beta = n_p/n$, where $n = n_c + n_p$, denote the fraction of vertices in the core. We assign the edges independently at random according to the following procedure. The edge probabilities for the core-core, core-periphery, and periphery-periphery pairs of vertices are given by the vector $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp})$, where $p_{cc} = k^2p$, $p_{cp} = kp$, and $p_{pp} = p$. In our simulations, we fix $n = 100$, $\beta = 0.5$, and $p = 0.25$, and we compute core-periphery structure for 10 instances of the above random-graph ensemble for each of the parameter values $k = 1.1, 1.2, \dots, 1.9, 2$. In Section 7, we include a detailed discussion of the results of these simulations, which we use to compare the performance of PATH-CORE with those of existing methods for determining core-periphery structure as well as with other new methods that we will introduce in the following sections. For now, to illustrate the effectiveness of the PATH-CORE algorithm, we show in Fig. 1 the PATH-CORE for all vertices for three different instances of the above block model. We use the parameter values

¹We will also sometimes use the notation $C = |V_C|$ for the size of the core set.

²In the Erdős-Rényi random graph model $G(n, p)$ on n vertices, recall that an edge is present between each pair of vertices independently with probability p .

$k = 1.3$ [which yields $\mathbf{p} = (0.4225, 0.325, 0.25)$], $k = 1.5$ [which yields $\mathbf{p} = (0.5625, 0.375, 0.25)$], and $k = 1.8$ [which yields $\mathbf{p} = (0.81, 0.45, 0.25)$].

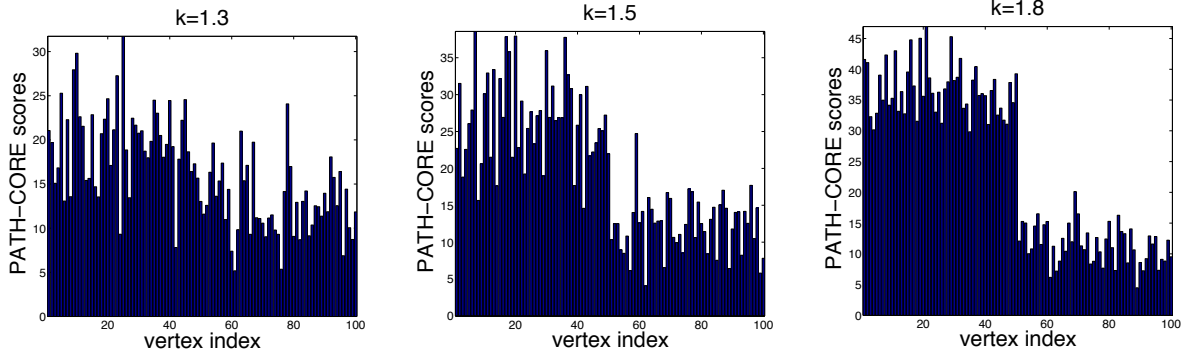


FIG. 1. PATH-CORE scores of all $n = 100$ vertices, for graphs drawn from three different random-graph ensembles in the family $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$. The vector $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp})$ gives the edge probabilities between a pair of core vertices (p_{cc}), a core vertex and a peripheral vertex (p_{cp}), and a pair of peripheral vertices (p_{pp}). These probabilities are $p_{cc} = k^2 p$, $p_{cp} = kp$, and $p_{pp} = p$, and we use the fixed value $p = 0.25$. The scalar k then parametrizes the ensemble. The values of k are (left) 1.3, (center) 1.5, and (right) 1.8. The first 50 vertices represent the planted core vertices, and the remaining 50 vertices are the planted peripheral vertices.

For each of the plots in Fig. 1, we place the core vertices in the first 50 positions on the horizontal axis, and we place the peripheral vertices in the remaining 50 positions. The vertical axis indicates the PATH-CORE score associated to each vertex. As expected, vertices in the core set have larger PATH-CORE scores than vertices in the periphery set. For $k = 1.3$ (left panel), the separation between core and peripheral vertices is not very clear. As we increase k , the separation becomes clearer, and $k = 1.8$ (right panel) exhibits a clear separation between core and peripheral vertices. As expected, larger differences between the edge probabilities $p_{cc} \geq p_{cp} \geq p_{pp}$ in the random-graph ensemble result in clearer separations between core and periphery sets.

For some problems, it is sufficient to have a coreness measure that reflects the probability that a vertex is a core or peripheral vertex. In such a scenario, we view such scores as akin to centrality values [69]. In other situations, however, it is desirable to obtain a classification of a network’s vertices as part of a core set or a periphery set. For the latter scenario, we let $\text{PATH-CORE}(i)$ denote the PATH-CORE score of vertex i , and we assume without loss of generality that $\text{PATH-CORE}(1) \geq \text{PATH-CORE}(2) \geq \dots \geq \text{PATH-CORE}(n-1) \geq \text{PATH-CORE}(n)$. Because the PATH-CORE score gives our calculation for the likelihood that a vertex is in the core set or periphery set (a high PATH-CORE suggests a core vertex), we are left with inferring what constitutes a good “cut” of PATH-CORE values to separate core vertices from peripheral ones. In other words, we need to determine a threshold ξ such that we classify i as a core vertex if $\text{PATH-CORE}(i) \geq \xi$ and we classify i as a peripheral vertex if $\text{PATH-CORE}(i) < \xi$.

If the size $n_c = \beta n$ of the core set is known, then the problem becomes significantly easier, as we can select the top n_c vertices with the largest PATH-CORE scores and classify them as core vertices. That is, we set $a = n_c = \beta n$. However, in most real (or realistic) scenarios, the size of the core is not known a priori, and it should thus be inferred from the graph G (or from the graph ensemble) and the distribution of the PATH-CORE scores. One possible heuristic approach to obtain such a separation is to sort the vector of PATH-CORE scores in decreasing order and to infer a by searching for large a jump in the sizes of the vector elements. That is, one can seek a “natural” separation between high and low PATH-CORE scores (if one exists). An alternative approach is to detect two clusters in the vector of PATH-CORE scores using a clustering algorithm (such as k -means clustering). The examples in Fig. 2 (which we generate from the random-graph ensemble $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with $p_{cc} = k^2 p$, $p_{cp} = kp$, and $p_{pp} = p$ for $k \in \{1.3, 1.5, 1.8\}$) illustrate this heuristic very well, as there exists a natural cut point that corresponds to a PATH-CORE score of approximately $a = 20$. This cut correctly assigns the first 50 vertices to the core set and the remaining 50 vertices to the periphery set.

Unfortunately, for “noisy” networks from this graph ensemble (and for many empirical networks), for which the edge probabilities p_{cc} , p_{cp} , and p_{pp} are not well-separated, the aforementioned heuristic procedure can yield unsatisfactory results, so a more systematic approach is desirable. In Section 4, we thus introduce the FIND-CUT algorithm, which maximizes an objective function for partitioning a network into a set of core vertices and a set of peripheral vertices. Using the vector of PATH-CORE scores as an input — or, indeed, using any other vector of scores that reflects the likelihood that each vertex belongs to the core set — we consider a large number of possible values of the vector to attempt to find an optimal separation of vertices into a core set and a periphery set

that maximizes the objective function in (6). See Section 4 for a discussion of this objective function and how we maximize it.

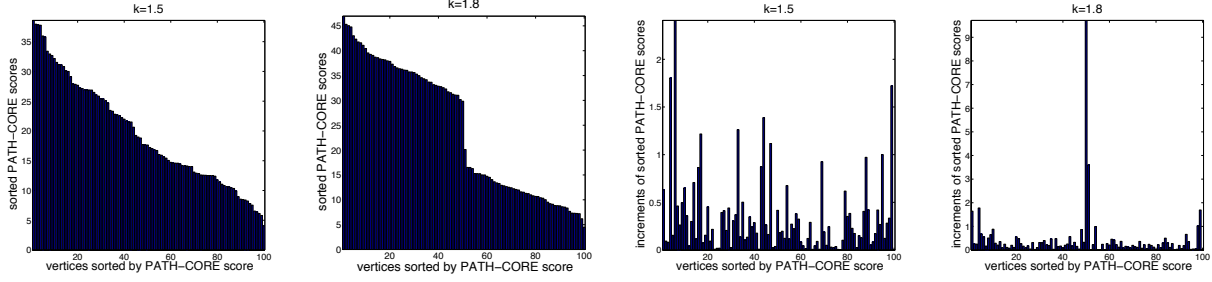


FIG. 2. PATH-CORE scores, sorted in decreasing order, for the random-graph ensemble $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with $p_{cc} = k^2 p$, $p_{cp} = kp$, and $p_{pp} = p$ and parameter values (left panel) $k = 1.5$ and (second panel) $k = 1.8$. When the core-periphery structure is sufficiently prominent, it is possible to separate the vertices by sorting the vector of PATH-CORE scores and inferring the threshold between core and peripheral vertices by considering the largest increment that occurs between two consecutive entries in the vector of sorted PATH-CORE scores. We show the result for $k = 1.5$ in the third panel and the result for $k = 1.8$ in the right panel. In the center and right panels of Fig. 1 and the rightmost two panels of the present figure, the largest PATH-CORE score of a peripheral vertex is approximately 20, whereas the lowest PATH-CORE score of a core vertex is approximately 30 (the difference of 10 is revealed by the peak in the right plot of this figure), and we obtain a clear discrete classification into a set of core vertices and a set of peripheral vertices.

We summarize the approach of PATH-CORE in Algorithms 5, 6, and 7 for the case of unweighted and undirected graphs. We present an algorithm to compute the PATH-CORE score for all vertices in a graph that runs in $\mathcal{O}(m^2)$ time, where we recall that $m = |E|$ is the number of edges in the graph. Intuitively, this is the best that one can achieve (even when computing a PATH-CORE score for just a single vertex), because one must separately consider each graph $G \setminus e$ for all $e \in E$, and finding shortest paths between two vertices has a complexity of $\Theta(m)$.

Let $G(V, E)$ be an unweighted graph without self-edges or multi-edges (i.e., it is a “simple” graph). Recall that we define the PATH-CORE score (2) of a vertex $i \in V$ as the sum over all adjacent vertex pairs in G of the fraction of shortest nontrivial paths containing i between each vertex pair in $V(G) \setminus i$. By “nontrivial,” we mean that the direct edge between those adjacent vertices does not count as a path. Our algorithm has some similarities to the algorithm presented in [15], and we follow some of the notation introduced therein. Let $d_G(j, i)$ be the “distance” between vertices j and i ; we define this distance as the minimum length of any path that connects j and i in G . Let $\sigma_{st}(i)$ be the number of shortest paths between s and t that contain i . Define the set of *predecessors* of a vertex i on shortest paths from s as

$$P_s(i) = \{j \in V : (j, i) \in E, d_G(s, i) = d_G(s, j) + 1\}.$$

We use the following observation: if i lies on a shortest path between s and t , then

$$\sigma_{st}(i) = \left(\sum_{k \in P_s(i)} \sigma_{si}(k) \right) \times \left(\sum_{l \in P_t(i)} \sigma_{it}(l) \right).$$

This will help to count the number of shortest paths on which a vertex lies without keeping track of the locations of these shortest paths. In the PATH-SCORE algorithm, $\sigma_s(i)$ is the number of paths between s and i of length $d_{G'}(s, i)$ if and only if i lies on a shortest path between s and t (i.e., if (s, t) is the edge that is currently removed), where G' is the graph $G \setminus (s, t)$. The algorithm records the distance between s and i in G' as $d_s(i)$. In Algorithm 5, we calculate PATH-CORE scores for every vertex. For this algorithm, we need to compute shortest-path distances between pairs of vertices (see Algorithm 6) and to count the numbers of shortest paths (see Algorithm 7). We detail the steps of our algorithms and their complexity results in Appendix 1.

3.2. SAMPLED-PATH-CORE and Computational Complexity. As we show in Appendix 1, our PATH-CORE algorithm has a temporal complexity of $\mathcal{O}(m^2)$, where m is the total number of edges in a network, and a spatial complexity of $\mathcal{O}(m)$. The temporal complexity can be prohibitive for large and dense graphs. One can, however, speed up the algorithm by sampling edges from a graph G via some random process and computing shortest paths only for pairs of adjacent vertices that use these sampled edges. In other words, we pick an edge $(i, j) \in E$ with some

probability α , compute the PATH-CORE scores of the vertices in $V \setminus (i, j)$. Note that we do not further sparsify the graph G , but we only choose to compute shortest paths between a subset of the adjacent vertices. This approach has the potential of significantly reducing the temporal complexity of the PATH-CORE algorithm, thereby making it amenable to computing core-periphery structure in very large networks. We refer to this modified algorithm as the SAMPLED-PATH-CORE algorithm. The drawback of such a sampling approach is that the variance of the PATH-CORE scores increases, and some vertices might cross over the threshold between core and peripheral vertices (and thus be classified incorrectly). An investigation of the trade-off between accuracy and computational efficiency of SAMPLED-PATH-CORE is beyond the scope of our paper, but it is an interesting direction for future research.

4. An Objective Function for Detecting core-periphery Structure. In this section, we introduce an objective function that is suitable for detecting core-periphery structure when there is exactly one core set of vertices and one periphery set. Using this objective function, we propose the FIND-CUT algorithm for the partitioning the vertex set V into core and periphery sets. As an input, FIND-CUT takes a vector of scores that reflect the likelihood that each vertex belongs in a network's core set (the probability of belonging to the core set is higher for larger scores), and it attempts to find an optimal separation that maximizes the proposed objective function. We believe that the construction of a suitable objective function brings three advantages. First, the subject of network community structure has benefited greatly from having objective functions to optimize [31, 67], and we expect similar boons for investigations of core-periphery structure. Second, it allows a local-refinement search after the initial algorithm has been applied (in the spirit of Kernighan-Lin vertex-swapping steps for community detection [57, 68] and gradient-descent refinement steps in non-convex optimization [55]). Finally, it allows one to compare distinct methods by comparing the corresponding value of the objective function. Nevertheless, one has to proceed cautiously: a value of an objective function need not provide a definitive answer, and it can sometimes even mislead practitioners [36].

Before introducing an objective function for studying core-periphery structure, we first revisit a well-known graph-partitioning problem to highlight the similarity between the two situations. MIN-CUT, an instance of a graph-partitioning problem, is concerned with dividing a graph into two (similarly-sized) subgraphs while minimizing the number of edges that are incident to vertices in both subgraphs. More generally, a large family of graph-partitioning problems seek to decompose a graph into k disjoint subgraphs (i.e., clusters) while minimizing the number of cut edges (i.e., edges with endpoints in different clusters). Given the number g of clusters, the g -way graph-partitioning problem searches for a partition V_1, \dots, V_g of the vertex set V that minimizes the number of cut edges

$$(3) \quad \text{Cut}(V_1, \dots, V_g) = \sum_{i=1}^g |E(V_i, \bar{V}_i)|,$$

where $\bar{X} = V \setminus X$ and the number of edges between $X \subset V$ and $Y \subset V$ is $|E(X, Y)| = \sum_{i \in X, j \in Y} A_{ij}$. However, it is well-known that trying to minimize $\text{Cut}(V_1, \dots, V_g)$ favors cutting off weakly-connected individual vertices from a graph and can thus lead to trivial partitions. To penalize clusters V_i of small size, Shi and Malik [74] suggested minimizing the normalized cut

$$(4) \quad \text{NCut}(V_1, \dots, V_g) = \sum_{i=1}^g \frac{\text{Cut}(V_i, \bar{V}_i)}{\text{SK}(V_i)},$$

where $\text{SK}(V_i) = \sum_{i \in V_i} d_i$ and d_i denotes the degree of vertex i in the original graph G .

A natural choice for an objective function to detect core-periphery structure is to maximize the number of edges between pairs of core vertices and also between core and peripheral vertices, while allowing as few edges as possible between pairs of peripheral vertices. In other words, our approach is complementary to that of the graph-cut objective function (3). However, instead of minimizing the number of cut edges across the core and periphery sets (i.e., across clusters), we maximize the connectivity between pairs of core vertices and between core and peripheral vertices while minimizing the connectivity between pairs of peripheral vertices. We thus want to maximize

$$(5) \quad \text{CP-connectivity}(V_C, V_P) = E(V_C, V_C) + E(V_C, V_P) - E(V_P, V_P).$$

Our aim is to find a partition $\{V_C, V_P\}$ of the vertex set V that maximizes $\text{CP-connectivity}(V_C, V_P)$, under the constraint that $|V_C|, |V_P| \geq b$, where b is the minimum number of core or peripheral vertices (hence, $n - b$ is the maximum number of core or peripheral vertices) to avoid a large imbalance between the sizes of the core and periphery sets. In other words, we seek a balanced partition, and a higher value of b indicates a smaller difference

between the sizes of the core and periphery sets. This constraint is required to avoid a trivial solution in which all of the vertices are placed in the core set. Furthermore, note that the objective function (5) has only one variable because of the constraint $E(V_C, V_C) + E(V_C, V_P) + E(V_P, V_P) = m$. In practice, we have found this approach to be rather unstable in the sense that (5) often attains its maximum at $|V_C| = b$ or $|V_P| = b$. It thereby leads to disproportionately-sized sets of core and peripheral vertices compared to the “ground truth” in problems with planted core–periphery structure [e.g., from the block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$]. This situation is analogous to the trivial solution that one obtains for unconstrained graph-partitioning problems. We have been able to alleviate this problem (though not completely) by incorporating a normalization term in the spirit of the normalized cut function (4). Instead of maximizing the number of edges between core vertices and between core and peripheral vertices while minimizing the number of edges between peripheral vertices, we choose to maximize the edge *density* among core vertices and between core and peripheral vertices while minimizing the edge density among peripheral vertices. Finally, we also add a term to the objective function that penalizes imbalances between the sizes of the core and periphery sets (or penalizes a deviation from the expected proportion of core vertices) if such information is available. The maximization of our new objective function is over the set of all possible partitions of the vertex set into two disjoint sets (the core set V_C and the periphery set V_P). The function is

$$(6) \quad \text{CP-density}(V_C, V_P) = \frac{|E(V_C, V_C)|}{\text{Vol}(V_C, V_C)} + \frac{|E(V_C, V_P)|}{\text{Vol}(V_C, V_P)} - \frac{|E(V_P, V_P)|}{\text{Vol}(V_P, V_P)} - \gamma \left| \frac{|V_C|}{n} - \beta \right|,$$

where

$$(7) \quad \text{Vol}(X, Y) = \begin{cases} |X||Y|, & \text{if } X \neq Y \\ \frac{1}{2}|X|(|X| - 1), & \text{if } X = Y \end{cases}$$

denotes the total possible number of connections between sets X and Y . In the penalty term, β denotes the prescribed fraction of core vertices in the graph (if it is known in advance), and γ tunes the sensitivity of the objective function to the size imbalance between the core and periphery sets. Note that β can either be prescribed in advance or construed as a parameter that guides the maximization towards a solution with a certain target size for the core set. For simplicity, we limit ourselves to the case $\gamma = 0$. That is, we assume no prior knowledge of the ratio between the number of core and peripheral vertices.³ For some of our experiments on synthetic graphs in Section 7, we compare the performance of our proposed algorithms both when β is known and when it is unknown.

Algorithm 1 FIND-CUT: Classifies the vertices of a graph G into a set V_C of core vertices and a set V_P of peripheral vertices based on a score associated to each vertex that reflects the likelihood that it is in the core.

Input: Vector of scores $\mathbf{s} = (s_1, \dots, s_n) \in \mathbb{R}^n$ associated to the n vertices of a graph.

- 1: Sort the entries of the vector \mathbf{s} in decreasing order. Assume without loss of generality that $s_1 \geq s_2 \geq \dots \geq s_{n-1} \geq s_n$.
- 2: Let $X_C = \{1, \dots, n_c\}$ and $Y_C = \{n_c + 1, \dots, n\}$ for all $n_c \in \{1, \dots, n\}$. Find the value n_c that maximizes the following objective function [which is similar to Eq. (6)]:

$$(8) \quad \Phi^* = \frac{1}{n} \left[\arg \max_{n_c \in \{b, \dots, n-b\}} \left(\frac{|E(X_C, X_C)|}{\text{Vol}(X_C, X_C)} + \frac{|E(X_C, Y_C)|}{\text{Vol}(X_C, Y_C)} - \frac{|E(Y_C, Y_C)|}{\text{Vol}(Y_C, Y_C)} \right) \right],$$

where b denotes a lower bound on the size of the core and periphery sets (which we use to avoid solutions with either a very small core set or a very small periphery set).

- 3: Define the core set $V_C = \{1, \dots, n_c\}$ and the periphery set $V_P = \{n_c + 1, \dots, n\}$
-

We summarize the FIND-CUT approach in Algorithm 1, and we remark that one can also add an iterative post-processing refinement step that are reminiscent of the gradient-descent algorithm [55] or of Kernighan–Lin vertex swaps [57, 68]. At each iteration, one can choose to move the vertex from the core set to the periphery set (or the other way around) that leads to the largest increase in the objective function (6). Alternatively, if one wishes to maintain the current size of the core and periphery sets, then one can choose to swap a pair of vertices from their assignments (of core or periphery) that leads to the largest increase in the objective function.

³In practice, however, we do implicitly assume a lower bound on the sizes of the core and periphery sets of vertices to ameliorate a “boundary effect” that yields solutions with a very small number of vertices in the core set or periphery sets. If one explicitly wants to allow the possibility of a small set of core or peripheral vertices, then one can set $b = 0$.

5. LOWRANK-CORE: Core-Periphery Detection via Low-Rank Matrix Approximation. Another approach for detecting core-periphery structure in a network is to interpret its adjacency matrix as a perturbation of a low-rank matrix. Consider, for instance, the block model

$$(9) \quad A_0 = \begin{bmatrix} \mathbf{1}_{n_c \times n_c} & \mathbf{1}_{n_c \times n_p} \\ \mathbf{1}_{n_p \times n_c} & \mathbf{0}_{n_p \times n_p} \end{bmatrix},$$

which assumes that core vertices are fully connected among themselves and with all vertices in the periphery set and that no edges exist between any pair of peripheral vertices. The block model in Eq. (9) corresponds to an idealized block model that Borgatti and Everett [12] employed in a discrete notion of core-periphery structure. The rank of the matrix A_0 is 2, as any 3×3 submatrix has at least two identical rows or columns. Consequently, $\det(A_0) = 0$. Alternatively, when the core and periphery sets have the same size, $n_c = n_p$ with $n = n_c + n_p$, one can write the matrix A_0 as the following tensor product of matrices:

$$(10) \quad \bar{A}_0 = \begin{bmatrix} \mathbf{1}_{n_c \times n_c} & \mathbf{1}_{n_c \times n_c} \\ \mathbf{1}_{n_c \times n_c} & \mathbf{0}_{n_c \times n_c} \end{bmatrix} = R \otimes \mathbf{1}_{n_c \times n_c}, \quad R = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

The eigenvalues of \bar{A}_0 are direct products of the eigenvalues of R and $\mathbf{1}_{n_c \times n_c}$. That is, they are

$$\left\{ \frac{1 - \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2} \right\} \otimes \{n_c, 0^{(n_c-1)}\} = \left\{ \left(n_c \frac{1 \pm \sqrt{5}}{2} \right)^{(2)}, 0^{(n-2)} \right\},$$

where a superscript denotes the multiplicity of an eigenvalue.

The simplistic block models in equations (9,10) assume that a network has only one core set and one periphery set. Consequently, the block-model matrix has a rank of 2. The matrix rank is higher for more complicated core-periphery block models. For example, the block model in Fig. 3 has a global community structure — there are $g = 4$ communities, which each correspond to a block in the block-diagonal matrix — and a local core-periphery structure (because each community has a core-periphery structure). As indicated in Ref. [69], one can also construe such a structure (by permuting the rows and columns of the matrix) as having a global core-periphery structure and a local community structure.

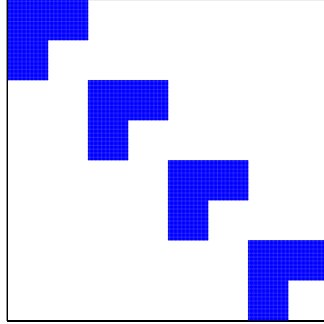


FIG. 3. A block model with $g = 4$ diagonal blocks that are each of the form of the block model in Eq. (9).

Let $B_g(A_0)$ denote a “hierarchical” ensemble of size $n \times n$ that is composed of g diagonal blocks that are of each of size $l \times l$ (thus, $n = lg$), where each diagonal block is of the form of the block model in Eq. (9). If we let λ_1, λ_2 denote the two nonzero eigenvalues of A_0 and let I_g denote the identity matrix of size g , then we can also write $B_g(A_0)$ as a tensor product of matrices:

$$(11) \quad B_g(A_0) = I_g \otimes A_0, \quad \text{with eigenvalues} \quad \{B_g(A_0)\} = \{1^g\} \otimes \{\lambda_1, \lambda_2, 0^{l-2}\} = \left\{ \lambda_1^{(g)}, \lambda_2^{(g)}, 0^{(n-2g)} \right\}.$$

Therefore, in the simplistic scenario in which each diagonal block has one core set and one periphery set (and thus has rank 2), the rank of $B_g(A_0)$ is $2g$.

Motivated by the low-rank structure of the above block-model networks, it is useful to consider the possibility of recovering a network’s unknown structure using a simple low-rank projection of its adjacency matrix. For the

remainder of this section, we focus on the simple core-periphery structure whose rank-2 block model is given by Eq. (9) (with one core set and one periphery set). In practice, we construe the adjacency matrix A of an observed graph G as a low-rank perturbation of the block model A_0 . In other words, we decompose A as

$$(12) \quad A = A_0 + W,$$

where W is a “noise matrix” whose entries $\{-1, 0, 1\}$ are determined by a mixture model [51] that involves block-model parameters. The entries of W are

$$(13) \quad W_{ij} = \begin{cases} -1, & \text{with probability } 1 - p_{cc} & (\text{i.e., if } i, j \in V_C), \\ -1, & \text{with probability } 1 - p_{cp} & (\text{i.e., if } i \in V_C \text{ and } j \in V_P), \\ 1, & \text{with probability } p_{pp} & (\text{i.e., if } i, j \in V_P), \\ 0, & \text{otherwise.} \end{cases}$$

Note that W is a random block-structured matrix with independent entries, and its expected value is the rank-2 matrix with entries

$$(14) \quad \mathbb{E}(W_{ij}) = \begin{cases} p_{cc} - 1, & \text{if } i, j \in V_C, \\ p_{cp} - 1, & \text{if } i \in V_C \text{ and } j \in V_P, \\ p_{pp}, & \text{if } i, j \in V_P. \end{cases}$$

To “denoise” the adjacency matrix A and recover the structure of the block model, we consider its top two eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2\}$, whose corresponding two largest (in magnitude) eigenvalues are $\{\lambda_1, \lambda_2\}$, and we compute the rank-2 approximation

$$(15) \quad \hat{A} = [\mathbf{v}_1 \quad \mathbf{v}_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix}.$$

As A more closely approximates the block model, which we can construe as a sort of “null model”, the spectral gap between the top two largest eigenvalues and the rest of the spectrum becomes larger (as illustrated by the plots in the second column of Fig. 4). In other words, as the amount of noise in (i.e., perturbation of) the network becomes smaller, the top two eigenvalues $\{\lambda_1, \lambda_2\}$ become closer to the eigenvalues $\left\{ \lambda_1 = n_c \left(\frac{1+\sqrt{5}}{2} \right), \lambda_2 = n_c \left(\frac{1-\sqrt{5}}{2} \right) \right\}$ of the block model.

To illustrate the effectiveness of our low-rank projection and that the projected matrix \hat{A} is a better approximation to the stochastic block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ than the initial adjacency matrix A , we consider two synthetically generated networks based on the block model. To do this, we use the edge probabilities $(p_{cc}, p_{cp}, p_{pp}) = (0.7, 0.7, 0.2)$ and $(p_{cc}, p_{cp}, p_{pp}) = (0.8, 0.6, 0.4)$. In the left column of Fig. 4, we show their corresponding adjacency matrices. The spectrum, which we show in the middle column, reveals the rank-2 structure of the networks. In the second example (which we show in the bottom row of the figure), the large amount of noise causes the second largest eigenvalue value to merge with the bulk of the spectrum.

We then use the denoised matrix \hat{A} to classify vertices as part of the core set or the periphery set by considering the degree (i.e., the row sums of \hat{A}) of each vertex. We binarize \hat{A} by setting its entries to 0 if they are less than or equal to 0.5 and setting them to 1 if they are larger than 0.5, and we denote the resulting binarized matrix by \hat{A}_t .⁴ In the right column of Fig. 4, we show the recovered matrix \hat{A}_t for our two example networks. Note in both examples that the denoised matrix \hat{A}_t resembles the core-periphery block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ much better than the initial adjacency matrix A . Finally, we compute the degree of each vertex in \hat{A}_t , and we call these degrees the LOWRANK-CORE scores of the vertices. We use the LOWRANK-CORE scores to classify vertices as core vertices or peripheral vertices. If one knows the fraction β of core vertices in a network, then we choose the top βn vertices with the largest LOWRANK-CORE score as the core vertices. Otherwise, we use the vector of LOWRANK-CORE scores as an input to the FIND-CUT algorithm that we introduced in Sec. 4. Although a theoretical analysis of the robustness to noise of our low-rank approximation for core-periphery detection is beyond the scope of the present paper, we expect that results from the matrix-perturbation literature, such as Weyl’s inequality and the Davis–Kahan $\sin(\Theta)$ -theorem [11], as well results on low-rank deformations of large random matrices [10] (analogous to the results of Féral and Pécché on the largest eigenvalue of rank-1 deformations of real, symmetric random matrices [30])

⁴Following the rank-2 projection, we observe in practice that all entries of \hat{G} lie in the interval $[0, 1]$. We have not explored the use of other thresholds besides 0.5 for binarizing \hat{A} .

could lead to theoretical results characterizing the sparsity and noise regimes for which the above-proposed rank-2 projection is successful at separating core and peripheral vertices. A possible first step in this direction would be to consider a simplified version of the graph ensemble $G(p_{cc}, p_{cp}, p_{pp}, n_p, n_c)$ by setting $p_{cc} = p_{cp} = 1 - p_{pp} = 1 - \eta$, where $\eta \in (0, 1)$.

Algorithm 2 LOWRANK-CORE: Detects core-periphery structure in a graph based on a rank-2 approximation.

Input: Let A denote the adjacency matrix of the simple graph $G = (V, E)$ with n vertices and m edges.

- 1: Compute $\{\lambda_1, \lambda_2\}$, the top two largest (in magnitude) eigenvalues of A , together with their corresponding eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2\}$.
 - 2: Compute \hat{A} , a rank-2 approximation of A , as shown in Eq. (15).
 - 3: Threshold the entries of \hat{A} at 0.5 (so that entries strictly above 0.5 are set to 1 and all other entries are set to 0), and let \hat{A}_t denote the resulting graph.
 - 4: Compute the LOWRANK-CORE scores as the vertex degrees of \hat{A}_t .
 - 5: If the fraction of core vertices β is known, identify the set of core vertices as the top βn vertices with the largest LOWRANK-CORE scores.
 - 6: If β is unknown, use the vector of LOWRANK-CORE scores as an input to the FIND-CUT algorithm in Algorithm 1.
-

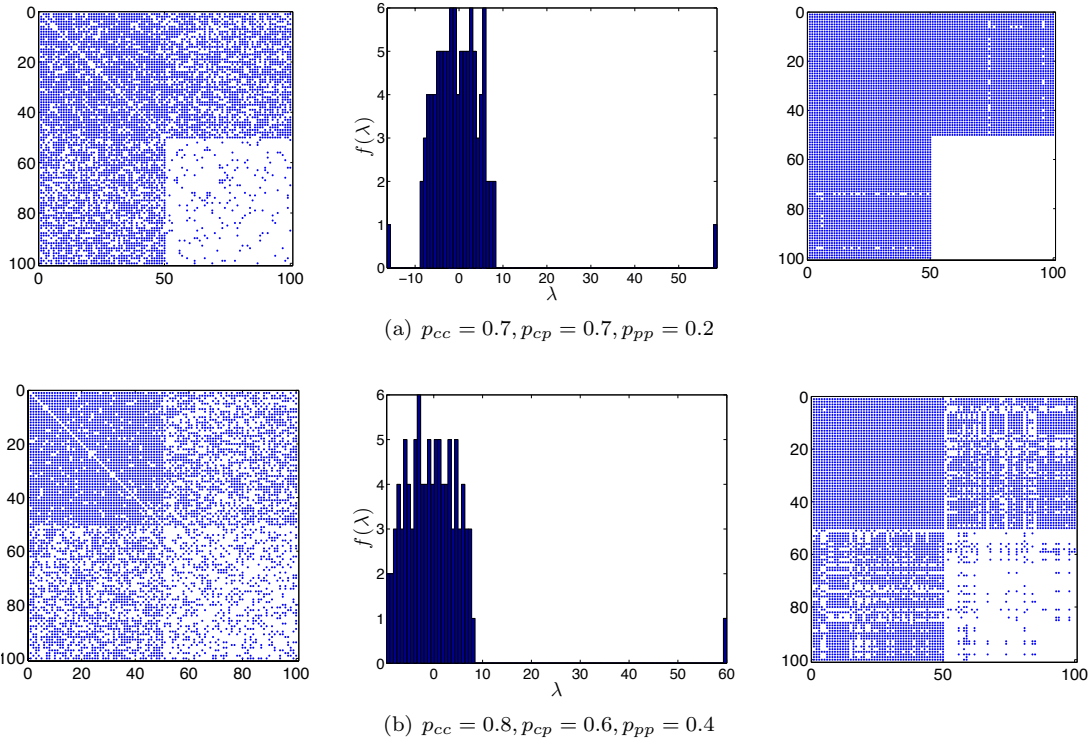


FIG. 4. (Left) Original adjacency matrices G from the stochastic block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with edge probabilities p_{cc} for edges between two core vertices, p_{cp} for edges between core vertices and peripheral vertices, and p_{pp} for edges between two peripheral vertices. (Middle) Histogram $f(\lambda)$ of the eigenvalues of the original adjacency matrices G . (Right) The matrices \hat{G}_t that we obtain after the rank-2 projection and thresholding.

6. LAP-CORE: Laplacian-Based Core-Periphery Detection. In this section, we explore the usefulness of using Laplacian eigenvectors for detecting core-periphery structure, and we remark that the methodology is applicable to both unweighted and weighted graphs. The *combinatorial Laplacian* matrix associated to the adjacency matrix G of a graph is $F = D - G$, where D is a diagonal matrix and D_{ii} denotes the degree of vertex i for an unweighted graph. For a weighted graph, D_{ii} denotes the sum of the weights associated to vertex i . The solutions of the generalized eigenvalue problem $F\mathbf{x} = \lambda D\mathbf{x}$ are related to the solutions of the eigenvalue problem $L\mathbf{x} = \lambda\mathbf{x}$,

where $L = D^{-1}G$ is often called the *random-walk Laplacian* of G . Using $L = I - D^{-1}F$, one can write the random-walk Laplacian in terms of the combinatorial Laplacian. Because L is a row-stochastic matrix, one can interpret it as a transition probability matrix of a Markov chain whose states are the vertices of G . In this interpretation, the matrix element L_{ij} denotes the transition probability that a random walker jumps from vertex i to vertex j in a single step. If the pair (λ, \mathbf{v}) is an (eigenvalue, eigenvector) solution to $L\mathbf{x} = \lambda\mathbf{x}$, then $(1 - \lambda, \mathbf{v})$ is a solution to $F\mathbf{x} = \lambda D\mathbf{x}$. The top⁵ eigenvectors of the random-walk Laplacian define the coarsest modes of variation or slowest modes of mixing in a graph, and they have a natural interpretation in terms of a random walk on the graph (and thus as a toy model of a conservative diffusion process). There exists a rich literature in the machine learning, data analysis, and image processing communities [9, 21, 22, 52, 73, 77] in successfully using such eigenvectors for tasks like clustering, ranking, image partitioning, and data visualization.

For core-periphery detection, it is useful to consider the bottom eigenvector of the associated random-walk Laplacian. Considering the block model in Eq. (9) or the generalized block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ (see the depiction in Table 1) with $p_{cc} \approx p_{cp} < p_{pp}$, the task of finding core-periphery structure in a given graph G amounts to trying to detect a dense connected component between the peripheral vertices in the complement graph \bar{G} (where the 0 non-diagonal entries of G become 1, and the 1 entries become 0), as such vertices have many non-edges between them in the original graph. In the above scenario when $p_{cc} \approx p_{cp} < p_{pp}$ and there exists a single densely-connected component in a given graph — such as for examples (a) and (b) in Fig. 5 — the eigenvector that corresponds to the second largest (in magnitude) eigenvalue of the associated random-walk Laplacian provides an accurate separation of the vertices in the dense component from the rest of the graph. The complement of the block-model graph has a periphery component of size n_p that is fully connected (i.e., it is K_{n_p} , the complete graph on n_p vertices), a core component without any edges between pairs of core vertices, and no edges between core and peripheral vertices. In practice, \bar{G} is a perturbed version of the above complement block model; that is, the peripheral vertices are very well-connected among themselves, and there are few core-core and core-periphery connections. The task becomes to identify a well-connected “community” of peripheral vertices. In other words, we have replaced the problem of identifying a core set and periphery set in G with the problem of finding the peripheral vertices in \bar{G} , for which we can use methods from the large set of available techniques for community detection [31, 67].

In many applications, the initial graph G is rather sparse, and the above approach thus has the drawback that the complement graph \bar{G} is a very dense graph, which significantly increases the time that is necessary for the computational task of identifying communities [16] (though we note that we only seek to identify a single dense subgraph rather than a graph’s entire community structure). As we discussed above, one way to find a dense subgraph of an initial graph is to use the first nontrivial eigenvalue (i.e., the second largest eigenvalue) of the random-walk Laplacian. In Fig. 5(a), we show an example of such a computation. In this case, we start with a block-model graph from $G(p_{cc} = 0.8, p_{cp} = 0.2, p_{pp} = 0.2, n_c, n_p)$, for which the first nontrivial eigenvalue (see the second column) clearly separates the planted dense subgraph from the rest of the network. In the eigenvector computation for the random-walk Laplacian, note that every iteration of the power method is linear in the number of edges in the graph, and the number of iterations is strictly greater than $\mathcal{O}(1)$ because it depends on the spectral gap. For sparse graphs G , the complement \bar{G} is a dense graph, which significantly increases the computational effort needed to find eigenvectors. Instead of working in the complement space, we turn our attention to the other end of the spectrum and consider the smallest eigenvalue of the random-walk Laplacian. Recall that all of the eigenvalues of the random-walk Laplacian are less than or equal to 1 in magnitude [19].

We now focus on the combinatorial Laplacian $F = D - G$. Let \bar{F} denote the combinatorial Laplacian associated to the graph \bar{G} . Note that $\bar{G} = J_n - G - I_n$, where J_n denotes the matrix of size $n \times n$ whose entries are all 1 and I_n is the $n \times n$ identity matrix. Additionally, $\bar{D} = (n - 1)I_n - D$. A well-known relationship [19] between the combinatorial Laplacian of a graph and that of its complement is given by

$$(16) \quad \bar{F} = \bar{D} - \bar{G} = (n - 1)I_n - D - (J_n - G - I_n) = nI_n - J_n - F.$$

If \mathbf{x} is an eigenvector of F (other than the trivial eigenvector $\mathbf{1}_n$) with $\mathbf{x} \perp \mathbf{1}_n$ (which implies that $J\mathbf{x} = \mathbf{0}$) and associated eigenvalue λ , then \mathbf{x} is also an eigenvector of \bar{F} (with associated eigenvalue $n - \lambda$). A result due to Kelmans [42–44] that connects the characteristic polynomial of the combinatorial Laplacian matrix of G to that of its complement implies that

$$(17) \quad \lambda_j(\bar{F}) = n - \lambda_{n+2-j}(F), \quad \text{for all } j \in \{2, \dots, n\}.$$

⁵The top eigenvectors of the random-walk Laplacian L are the eigenvectors that correspond to the largest eigenvalues of L . That is, these are the eigenvalues closest to $\lambda_1 = 1$, the largest eigenvalue of L . The bottom eigenvectors of L correspond to the smallest eigenvalues of L . Note that the eigenvalues $\lambda_1 = 1 \leq \lambda_2 \leq \dots \leq \lambda_n$ of L satisfy $|\lambda_i| \leq 1$ for all $i \in \{1, \dots, n\}$.

Equation (17) relates the eigenvalues of G to those of \bar{G} . In other words, the spectrum exhibits a certain symmetry, and questions regarding $\lambda_{n+2-j}(F)$ of a graph are equivalent to questions about $\lambda_j(\bar{F})$ of its complement. Furthermore, keeping in mind the usefulness of the second largest eigenvector of the combinatorial Laplacian, we stress that questions involving $\lambda_2(\bar{F})$ (i.e., the case $j = 2$) are equivalent to questions involving $\lambda_n(F)$.

In practice, none of the eigenvectors of the combinatorial Laplacian are able to distinguish a coherent core set and periphery set in a graph (or a single community in the graph’s complement). We calculate the top and bottom eigenvectors (and intermediate ones) of the combinatorial Laplacian and find that none of them captures the distinction between core and periphery sets. Instead, we are able to effectively separate core and periphery sets if we use the random-walk Laplacian L with the goal of identifying a dense subgraph in the complement graph \bar{G} . To do this, we calculate the second eigenvector $\bar{\mathbf{v}}_2$ of its associated Laplacian \bar{L} . Motivated by the analogy in the beginning of this section and the interplay between the bottom eigenvalues of a graph and the top eigenvalues (and their associated eigenvectors) of the graph’s complement in the case of the combinatorial Laplacians F and \bar{F} , we propose to use the bottom eigenvalue (and its associated eigenvector) of the random-walk Laplacian L associated with our initial graph G . The downside of working with the random-walk Laplacian L is that (to the best of our knowledge) there does not exist a statement similar to Eq. (17). In Appendix 1, explain that such a symmetry exists for the random-walk Laplacian only under certain conditions.

In Algorithms 3 and 4, we summarize the main steps of two viable algorithms for core–periphery detection using the random-walk Laplacian of a graph. The only difference between Algorithms 3 and 4 is as follows. The former uses the entries of \mathbf{v}_n (the bottom eigenvector that corresponds to the smallest algebraic⁶ eigenvalue as an input to the FIND-CUT algorithm to infer an optimal separation of the vertices into core and periphery sets by maximizing the objective function (6). By contrast, in Algorithm 4, the same bottom eigenvector \mathbf{v}_n of the random-walk Laplacian provides an implicit threshold (i.e., the value 0), and one is able to classify each vertex as part of a core set or a periphery set by considering the sign of each entry. To choose a global sign, we multiple by -1 if necessary to maximize the objective function (6) and ensure that the positive entries correspond to core vertices. (If \mathbf{v}_n is an eigenvector of L , then so is $-\mathbf{v}_n$.)

Algorithm 3 LAP-CORE: Detects core–periphery structure in a graph using a core score that is based on the eigenvector that corresponds to the smallest nonzero eigenvalue of the associated random-walk graph Laplacian.

Input: Let A denote the adjacency matrix of the simple graph $G = (V, E)$ with n vertices and m edges.

- 1: Compute the random-walk Laplacian $L = D^{-1}A$, where D is a diagonal matrix with elements $D_{ii} = \sum_{j=1}^n A_{ij}$ given by the strength (i.e., the sum of weights of the edges incident to the vertex) of vertex i for each i .
 - 2: Compute λ_n , which denotes the smallest nonzero eigenvalue of L , and its corresponding eigenvector \mathbf{v}_n . The eigenvector components give the LAP-CORE scores of the vertices.
 - 3: If β is known, identify the set of core vertices as the top βn vertices with largest LAP-CORE scores.
 - 4: If β is unknown, use the vector of LAP-CORE scores as an input to the FIND-CUT algorithm.
-

Algorithm 4 LAPSGN-CORE: Detects core–periphery structure in a graph using the signs of the components of the eigenvector that corresponds to the smallest nonzero eigenvalue of the associated random-walk graph Laplacian.

Input: Let A denote the adjacency matrix of the simple graph $G = (V, E)$ with n vertices and m edges.

- 1: Compute the random-walk Laplacian $L = D^{-1}A$.
 - 2: Compute λ_n , which is the smallest nonzero eigenvalue of L , and its corresponding eigenvector \mathbf{v}_n . The eigenvector components give the LAP-CORE scores of the vertices.
 - 3: Set $z_i = \text{sign}(\mathbf{v}_n(i))$ for $i \in \{1, \dots, n\}$. Because the eigenvector \mathbf{v}_n is determined up to a global sign change, do the following:
 - 4: Let vertex $u_i \in V_C$ if $z_i \geq 0$, and otherwise let $u_i \in V_P$. Let η_1 denote the resulting value of the objective function (6).
 - 5: Let vertex $u_i \in V_C$ if $z_i \leq 0$, and otherwise let $u_i \in V_P$. Let η_2 denote the resulting value of the objective function (6).
 - 6: If $\eta_1 > \eta_2$, let the final solution be $u_i \in V_C$ if $z_i \geq 0$; otherwise, let $u_i \in V_P$.
 - 7: If $\eta_2 > \eta_1$, let the final solution be $u_i \in V_C$ if $z_i \leq 0$; otherwise, let $u_i \in V_P$.⁷
-

⁶Because all of the random-walk Laplacian eigenvalues are real and no larger than 1 in magnitude, the smallest algebraic eigenvalue corresponds to the smallest real eigenvalue.

To illustrate the above interplay between the top and bottom parts of the spectrum of the random-walk Laplacian matrix, we consider the stochastic block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$, where we fix the core–core interaction probability $p_{cc} = 0.8$ and the periphery–periphery interaction probability $p_{pp} = 0.3$, but we vary the core–periphery interaction probability $p_{cp} \in [0.3, 0.7]$ in increments of 0.1. The goal of these numerical experiments, whose results we show in Fig. 5, is to demonstrate the ability of the bottom eigenvector \mathbf{v}_n of L to reveal a core–periphery separation when one exists (in contrast to the second largest eigenvector \mathbf{v}_2). To help visualize our results, we also employ a two-dimensional representation of the network vertices in which the core vertices (i.e., the vertices in the set V_C) are concentrated within a disc centered at the origin and the peripheral vertices (i.e., the vertices in the set V_P) lie on a circular ring around the core vertices. In Fig. 6, we plot the spectrum of the random-walk Laplacian associated to each of the p_{cp} values in the above experiment. Note that we disregard the trivial eigenvector $\mathbf{v}_1 = \mathbf{1}_n$ that corresponds to the trivial eigenvalue $\lambda_1 = 1$ of L .

For small values of p_{cp} (e.g., $p_{cp} = 0.3$ or $p_{cp} = 0.4$), the network does not exhibit core–periphery structure. Instead, it has a single community that is represented by the densely connected graph of vertices in the set V_C . As expected, the eigenvector \mathbf{v}_2 is able to highlight the separation between the V_C and V_P vertices very well, whereas the bottom eigenvector \mathbf{v}_n is not particularly helpful. For $p_{cp} = 0.5$, neither of the two eigenvectors above are able to capture the separation between V_C and V_P . However, as p_{cp} increases to $p_{cp} = 0.6$ and $p_{cp} = 0.7$ — such that we are closer to the idealized block model in (9) — there now exists a densely-connected subgraph of V_P in the complement graph \bar{G} . Instead of using the top nontrivial eigenvector $\bar{\mathbf{v}}_2$ of \bar{L} , we use the eigenvector \mathbf{v}_n that corresponds to the smallest eigenvalue λ_n of G , as this eigenvector is able to highlight core–periphery structure in G . In Fig. 6(a), we show that there is a clear separation between λ_2 and the bulk of the spectrum. Similarly, Fig. 6(e) illustrates a clear separation between λ_n and the bulk of the spectrum. For intermediate values of p_{cp} , such a spectral gap is significantly smaller or even nonexistent.

7. Numerical Experiments. In this section, we conduct a series of numerical experiments to compare different methods for detecting core–periphery structure and to assess the robustness of our methods to perturbations of a network. In Section 7.1, we examine synthetic networks with a global community structure and local core–periphery structure. In Section 7.2, we apply our various methods for detecting core–periphery structure to several empirical data sets. In Appendix 3, we examine networks with “planted” high-degree vertices in the periphery. Throughout this section, we refer to DEGREE-CORE as the method of detecting core–periphery structure by simply computing the vertex degrees and then applying the FIND-CUT method. Recall that this is for the situation of assuming a lower bound on the sizes of the core and periphery sets by having knowledge of the “boundary” sizes.

η	DEGREE-CORE	PATH-CORE	LOWRANK-CORE	LAP-CORE	LAPSGN-CORE	CORE-SCORE
$ V_C \geq 20, V_P \geq 20$	38 + 0	38 + 0	38 + 0	0 + 30	6 + 7	41 + 1
$ V_C \geq 50, V_P \geq 50$	3 + 5	3 + 5	26 + 1	0 + 22	6 + 7	27 + 2
Known $\beta = 0.5$	4 + 4	4 + 4	4 + 4	7 + 7	6 + 7	11 + 11

TABLE 2

Results for the example in Fig. 7 for (top two rows) unknown β value and (third row) known β value. When the value of β is unknown, we assume upper and lower bounds on the sizes of the core and periphery sets. Note the LAPSGN-CORE method does not need information about the sizes of the core and periphery sets, as it assigns vertices to such sets based only on the sign of the bottom eigenvector of the associated random-walk Laplacian. Each entry in the table counts the number of errors and is of the form $y_1 + y_2$, where y_1 indicates the number of core vertices that we label as peripheral vertices, and y_2 indicates the number of peripheral vertices that we label as core vertices.

As we illustrate in Fig. 7, the LAPSGN-CORE method yields the same results whether or not we impose lower bounds on the sizes of the core and periphery sets (e.g., to 10 or to 25 vertices), as it does not rely on information about the size of the core set. As we discussed in Section 6, it depends only on the sign of the entries of the top eigenvector of L . All of the other methods that we examine suffer from a “boundary effect,” as the FIND-CUT algorithm finds a global optimum at (or very close to) the boundary of the search interval. When β is known, we are planting core and periphery sets of known sizes, so we can examine the number of false-positive errors (i.e., vertices incorrectly assigned to the core set) and false-negative errors (i.e., vertices incorrectly assigned to the periphery set) for the various methods for detecting core–periphery structure. If we enforce a minimum size of 20 for the core and periphery sets, we find that LAPSGN-CORE is the only method that yields satisfactory results from this perspective, because all other methods find a maximum of the objective function that lies close to the boundary. When we increase the lower bound of the core and periphery sets from 20 to 50, the DEGREE-CORE and PATH-CORE methods yield very good results (in terms of the numbers of false positives and false negatives),

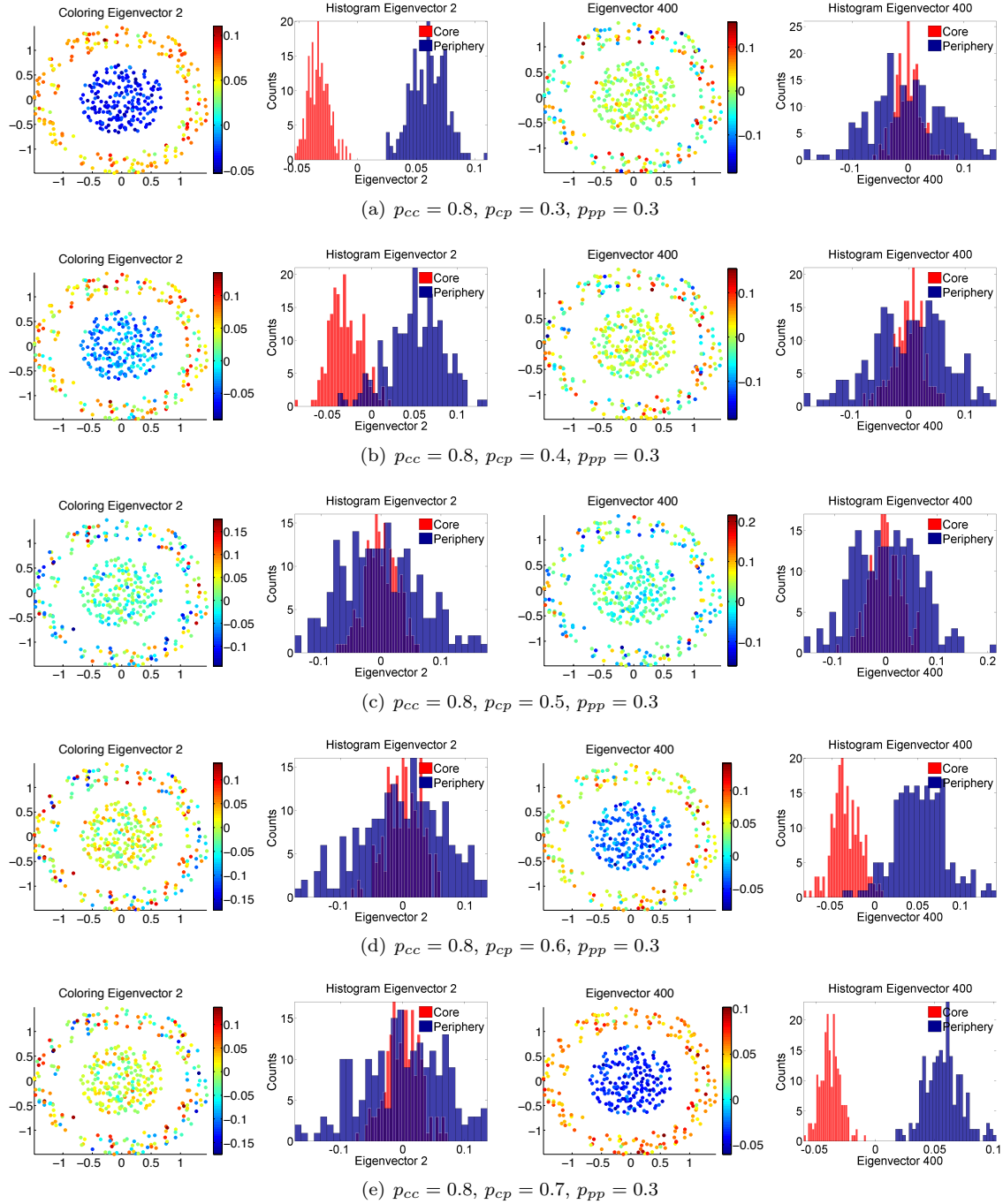


FIG. 5. [Color] Our simulations illustrate the interplay between the top and bottom parts of the spectrum of the random-walk Laplacian matrix L as a network transitions from a block model with block-diagonal “community structure” to a block model with core-periphery structure. Each row uses the stochastic block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with $n = 400$ vertices (with 200 core and 200 peripheral vertices) with a fixed core-core interaction probability $p_{cc} = 0.8$, a fixed periphery-periphery interaction probability $p_{pp} = 0.3$, and a varying core-periphery interaction probability $p_{cp} \in [0.3, 0.7]$. We vary p_{cp} in increments of 0.1, so the top row is for $p_{cp} = 0.3$, the second row is for $p_{cp} = 0.4$, and so on. The first and third columns give a coloring of a two-dimensional visualization of the graph vertices: the core vertices are contained in a disc that is centered at the origin, and the peripheral vertices lie on a ring around the core vertices. The second and fourth columns, respectively, show histograms of the entries of the eigenvectors \mathbf{v}_2 and \mathbf{v}_{400} . These eigenvectors correspond, respectively, to the largest (nontrivial) and smallest eigenvalues of the associated random-walk Laplacian matrix. The red color indicates core vertices, and the blue color indicates peripheral vertices. In Fig. 6, we plot the spectrum associated to each of the above six networks.

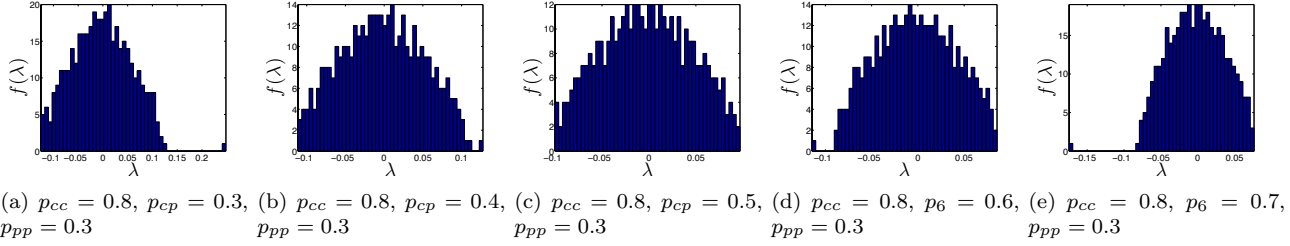


FIG. 6. Spectra $f(\lambda)$ of random-walk Laplacian matrices' eigenvalues λ for several instances of the stochastic block model $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$. In Fig. 5, we plotted histograms of the eigenvectors that correspond to the smallest and second largest eigenvalues for these matrices.

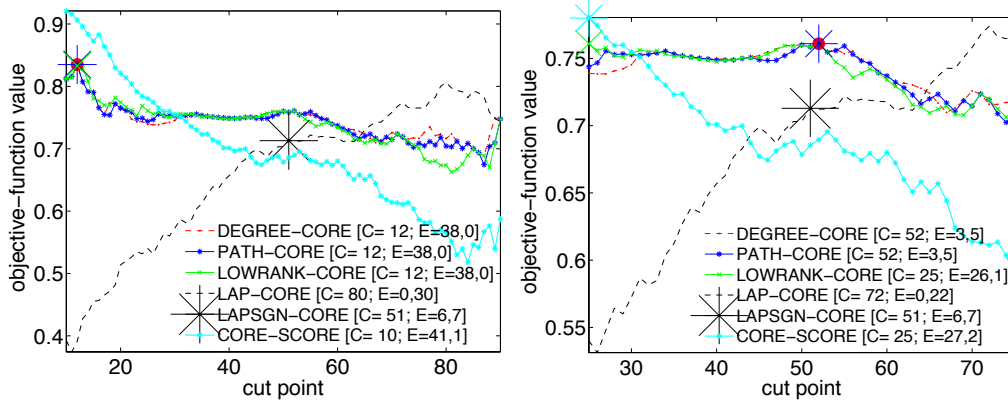


FIG. 7. Comparison of methods for detecting core-periphery structure for a graph from the ensemble $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with $n = 100$ vertices (and, in particular, $n_c = 50$ core vertices and $n_p = 50$ peripheral vertices) and edge probabilities $(p_{cc}, p_{cp}, p_{pp}) = (0.5, 0.5, 0.27)$ for the objective function in Eq. (8). We assume a minimum size for the core and periphery sets of at least (left) 10 vertices and (right) 25 vertices. We mark the cut points that maximize the objective function in Eq. (8) on the curves with a large filled circle for LAPSGN-CORE and other symbols for the other methods. The cut point refers to the number of core vertices. In the legends, C denotes the size of the core set that maximizes the objective function Eq. (8), and $\mathbf{E} = (y_1, y_2)$ denotes the corresponding 2-vector of errors. The first component of \mathbf{E} indicates the number of core vertices that we label as peripheral vertices, and the second indicates the number of peripheral vertices that we label as core vertices.

followed by LAPSGN-CORE, LAP-CORE, LOWRANK-CORE, and CORE-SCORE. When the fraction of vertices that belong to the core is known, then DEGREE-CORE, PATH-CORE, and LOWRANK-CORE again yield the best results, followed by LAPSGN-CORE, LAP-CORE, and CORE-SCORE.

Again evaluating the methods in terms of the number of false positives and false negatives, one can increase the accuracy of the methods to detect core-periphery structure by considering other local maxima of the objective function (6), especially if one is searching further away from the boundary. However, for these examples, the LAPSGN-CORE and CORE-SCORE methods still yield unsatisfactorily results even when considering additional local minima. Interestingly, their objective functions are monotonic (increasing for the former and decreasing for the latter) with respect to the vector of sorted scores. After assigning vertices to a core set or peripheral set using any of the methods above, one can also add a post-processing step in the spirit of either the gradient-descent refinement step in non-convex optimization [55] or Kernighan–Lin vertex swaps in community detection [57, 68]

The critical eye may object that a separation based on vertex degree yields results that are as good as the other best-performing methods. However, the recent work of [85] demonstrated that DEGREE-CORE separation is suboptimal for certain types of networks, although in the presence of a weak core-periphery structure, no algorithm can do much better than DEGREE-CORE. When a network's core and periphery are separated very strongly, examining vertex degree also appears to be reasonable. However, for pronounced core-periphery structure that is neither too weak nor too strong (i.e., in the most interesting scenario for applications), one needs to use methods that are more sophisticated than simply considering vertex degrees. Reference [69] also includes a salient discussion of examining a network's core-periphery structure simply by computing vertex degrees. To illustrate the sensitivity of the DEGREE-CORE method to the presence of high-degree periphery vertices, we perform two numerical experiments

where we purposely plant high-degree vertices in the periphery set (see Appendix 3). In these experiments, the LAPSGN-CORE method achieves the lowest number of errors, whereas DEGREE-CORE is one of the worst performers. In addition, one can see from Table 4, which gives the Pearson and Spearman correlation coefficients for various coreness measures, the results of our proposed methods are only reasonably correlated with DEGREE-CORE, and they can thus return solutions differ significantly from naive separation based on vertex degree. From the perspective of applications, we note the work of Kitsak et al. [45] on the identification of influential spreaders in networks. Kitsak et al. argued that the position of a vertex relative to the organization of a network determines its spreading influence to a larger extent than any local property (e.g., degree) of a vertex. Their findings also suggest that a network’s core vertices (as measured by being in the k -core of a network with high k) are much better spreaders of information than vertices with merely high degree. Moreover, very recent followup work has suggested that many core spreaders need not have large degrees [53], further highlighting the substantive difference between core vertices and high-degree vertices.

7.1. A Family of Synthetic Networks. In this section, we detail our numerical results when applying our methods to a family of synthetic networks with a planted core–periphery structure. We again examine the performance of the methods with respect to how many core and peripheral vertices they classify correctly.

We use variants of the random-graph ensemble that was introduced in [69]. Let $C_1(n, \beta, p, \kappa)$ denote a family of networks with the following properties: n is the number of vertices, β is the fraction of vertices in the core, and the edge probabilities for core–core, core–periphery, and periphery–periphery connections are $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp})$, where $p_{cc} = \kappa^2 p$, $p_{cp} = \kappa p$, and $p_{pp} = p$. Let $C_2(n, \beta, p, \kappa)$ denote a family of networks, from a slight modification of the above model, in which the edge probabilities are now given by $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp})$, where $p_{cc} = \kappa^2 p$, $p_{cp} = \kappa p$, and $p_{pp} = \kappa p$. In our simulations, we fix $n = 100$, $\beta = 0.5$, and $p = 0.25$, and we examine core–periphery structure using each of the proposed methods. We average our results over 100 different instantiations of the above graph ensembles for each of the parameter values $\kappa = 1.1, 1.2, \dots, 2$. We also compare our results with the CORE-SCORE algorithm introduced in [69], and we remark that the results of CORE-SCORE correspond to only a single instantiation of the above ensembles.⁸

In Fig. 8, we examine the ensemble $C_1(n, \beta, p, \kappa)$ and find that PATH-CORE, DEGREE-CORE, LOWRANK-CORE, and CORE-SCORE yield similar results. When β is unknown, we find that DEGREE-CORE, PATH-CORE, LOWRANK-CORE, and CORE-SCORE yield similar results to each other. However, when β is known (i.e., when we assume a lower bound on the sizes of the core and periphery sets), we find that DEGREE-CORE and LOWRANK-CORE tend to perform slightly better than CORE-SCORE and PATH-CORE. As expected, the aggregate performance of the various algorithms improves significantly when we assume knowledge of β . Unfortunately, in both scenarios, the two Laplacian-based methods yield very poor results. Recall that LAPSGN-CORE yields exactly the same results both with and without knowledge of β , so we only show it in the plots without knowledge of β .

In Fig. 9, we plot our numerical results for the ensemble $C_2(n, \beta, p, \kappa)$. When β is unknown, DEGREE-CORE, PATH-CORE, LOWRANK-CORE, and CORE-SCORE again yield similar results. When we assume that β is known, we find that CORE-SCORE, LOWRANK-CORE, DEGREE-CORE still perform similarly to each other, and they all do slightly better than PATH-CORE. The Laplacian-based methods again perform very poorly, though LAP-CORE does slightly better than LAPSGN-CORE when β is unknown.

In Fig. 10, we consider a graph with a core–periphery structure from a random-graph ensemble with edge probabilities $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp}) = (\kappa, \kappa, 1 - \kappa)$ for different values of κ . The common feature of this set of experiments — both when the boundary size β is known and when it is unknown — is that DEGREE-CORE, LOWRANK-CORE, and PATH-CORE give the best results, whereas CORE-SCORE consistently comes in last place (except for doing somewhat better than the Laplacian-based methods for small values of x) in terms of accuracy. In Fig. 11, we consider the values of the objective function (8), averaged over 100 runs, that we obtain using the different partitions of a network’s vertices into core and periphery sets as we sweep along the sorted scores that we compute using each of the methods (except CORE-SCORE, which we omit for computational reasons). In Fig. 12, we compare the actual values of the objective function for a single experiment across all methods (including CORE-SCORE) as we vary the parameter x . We also show the evolution of the value of the objective function as we sweep through the vector of scores from each method.

In Fig. 13, we compare the computation times (in seconds and on a \log_{10} scale) for all of the methods that we examine. The computers that we use for this comparison have 12 CPU cores [Intel(R) Xeon(R) X5650 @ 2.67GHz] and have 48 GB RAM. The most computationally expensive method is CORE-SCORE, which is 1–2

⁸The inefficient running time of the CORE-SCORE algorithm renders it infeasible to average over 100 different instantiations of a graph ensemble.

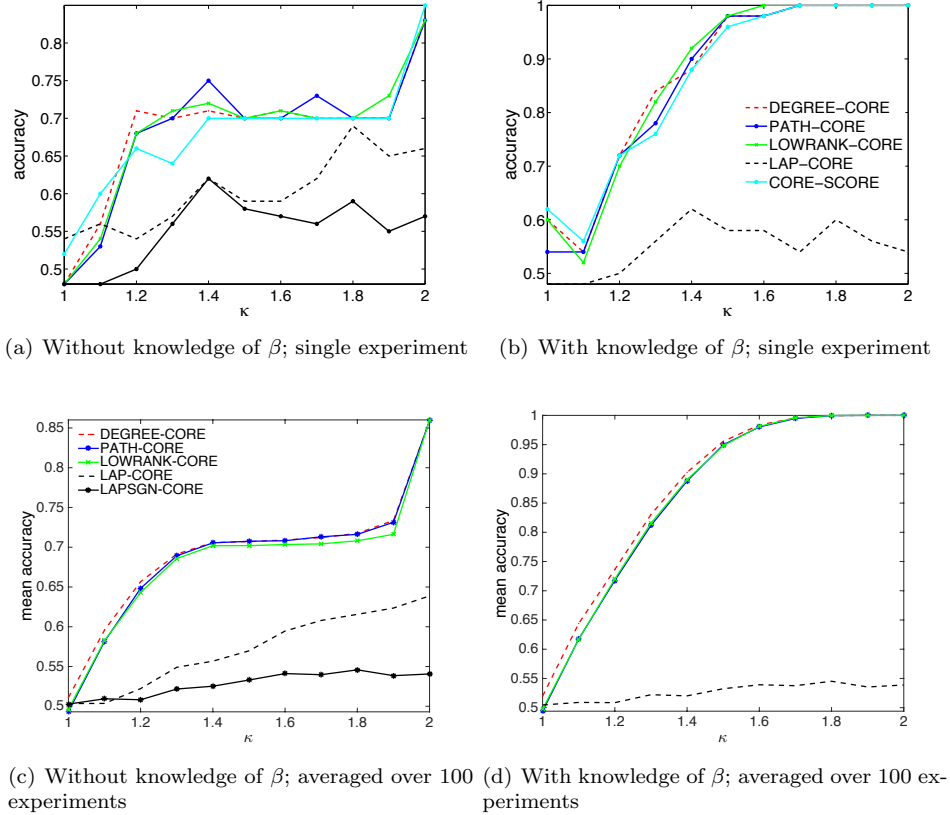


FIG. 8. Comparison of methods for core-periphery detection using the graph ensemble $C_1(n, \beta, p, \kappa)$ with $n = 100$, $\beta = 0.5$, $p = 0.25$, and edge probabilities $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp})$, where $p_{cc} = \kappa^2 p$, $p_{cp} = \kappa p$, and $p_{pp} = p$. We vary $\kappa \in [1, 2]$ in increments of 0.1. The top plots illustrate our results for all methods on a single graph from $C_1(n, \beta, p, \kappa)$, and the bottom plots give results averaged over 100 different graphs from the ensemble for all methods except CORE-SCORE. The left plots do not use information about the size (β) of the core, as they rely only on the objective function that one maximizes; the right plots explicitly use knowledge of β . The colors and symbols in the legend in (c) are also apply to (a), and the colors and symbols in the legend in (b) also apply to (d).

orders-of-magnitude slower than PATH-CORE, which is in turn 3–4 orders of magnitude slower than the spectral LOWRANK-CORE and LAP-SCORE methods (which have very similar computation times). Finally, as expected, the trivial DEGREE-CORE method has the fastest computation times.

7.2. Application to Empirical Data. In a recent publication, several of us applied the PATH-CORE and CORE-SCORE algorithms for detecting core-periphery structure in a variety of real-world networks. In the present paper, we use our various methods on a few empirical data sets.

We consider four examples of social networks using all of the methods that we have discussed for detecting core-periphery structure. The first two graphs are publicly-available networks of network scientists from 2006 (NNS2006) [57] and 2010 (NNS2010) [28] with 379 and 552 vertices, respectively, in their largest connected components (LCCs). Reference [69] considered core-periphery structure in both of these networks. The vertices are scholars (predominantly from physics) who study network science, and the weight of each (undirected) edge represents the strength of a coauthorship relationship. (See the original references for additional discussion of these networks and for more details about the weights, which are not necessarily defined in the same way in the two networks.) The other two networks are two universities (Caltech and Reed College) from the Facebook100 data set [80, 81], which consists of a single-time snapshot from the online social network Facebook in September 2005 for each of 100 universities in the United States. Caltech has 762 vertices in its LCC, and Reed has 962 vertices in its LCC.

In Figs. 14 (for the networks of network scientists) and 15 (for the Facebook networks), we present the objective-function values (8) for each method for detecting core-periphery structure. In Table 4 in the Appendix 4, we compare the Pearson and Spearman correlations between the coreness values of the methods for these empirical networks.

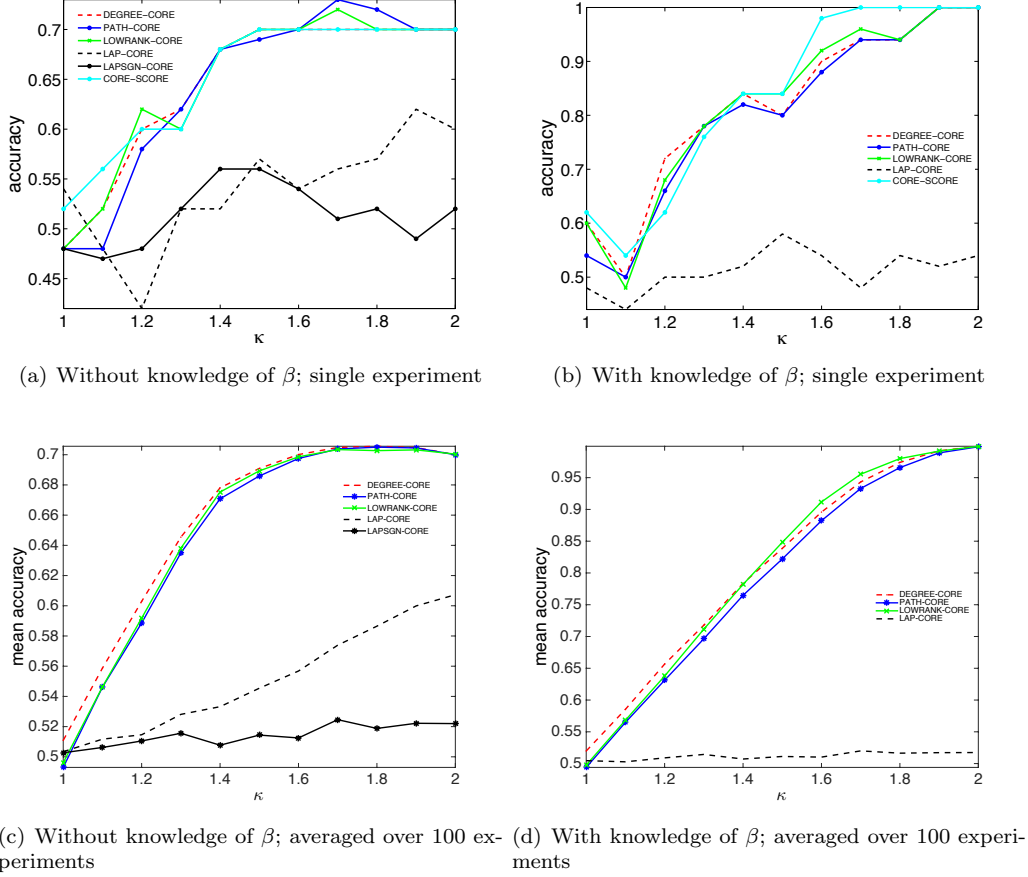


FIG. 9. Comparison of methods for detecting core-periphery structure for the graph ensemble $C_2(n, \beta, p, \kappa)$ with $n = 100$, $\beta = 0.5$, $p = 0.25$, and edge probabilities $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp})$, where $p_{cc} = \kappa^2 p$, $p_{cp} = \kappa p$, and $p_{pp} = \kappa p$. We vary $\kappa \in [1, 2]$ in increments of 0.1.

For these networks, we find that the values of DEGREE-CORE, CORE-SCORE, PATH-CORE, and LOWRANK-CORE are usually strongly correlated to each other, whereas the LAP-CORE values are very different. We find similar results when we use a similarity measure to compare partitions into a core set and periphery set from maximizing the objective function (8). We define the similarity fraction between two measures as

$$(18) \quad S_{\text{frac}} = \frac{w_1}{w_1 + w_0},$$

where w_1 is the number of vertices classified as core vertices in both measures or as peripheral vertices in both measures, and w_0 is the number of vertices that are classified differently in the two measures. (Thus, $w_0 + w_1 = n$ is the total number of vertices.) One can also observe that the two networks of network scientists are similar to each other and that the two Facebook networks are similar to each other in terms of their correlations and core-periphery partitions. See Table 4 in Appendix 4, and also see Figs. 14 and 15. For instance, the core-periphery separation points of LAP-CORE and LAPSGN-CORE yield much closer S_{frac} values for Facebook networks than for the networks of network scientists.

8. Summary and Discussion. We introduced several new methods for detecting core-periphery structure in graphs, and we compared these methods to existing ones using both synthetic and empirical networks. Our approach based on transportation relies on computing shortest paths in a graph between a pair of adjacent vertices after temporarily removing the edge between the two vertices. Another approach, which is motivated by the existence of a low-rank structure in networks that exhibit core-periphery structure, relies on a low-rank approximation of the adjacency matrix of a graph. We also introduced two methods that rely on the bottom eigenvector of the random-walk Laplacian associated to a graph. Finally, we introduced an objective function that helps in the classification of vertices into core and peripheral vertices, and we showed how one can use this objective function after obtaining

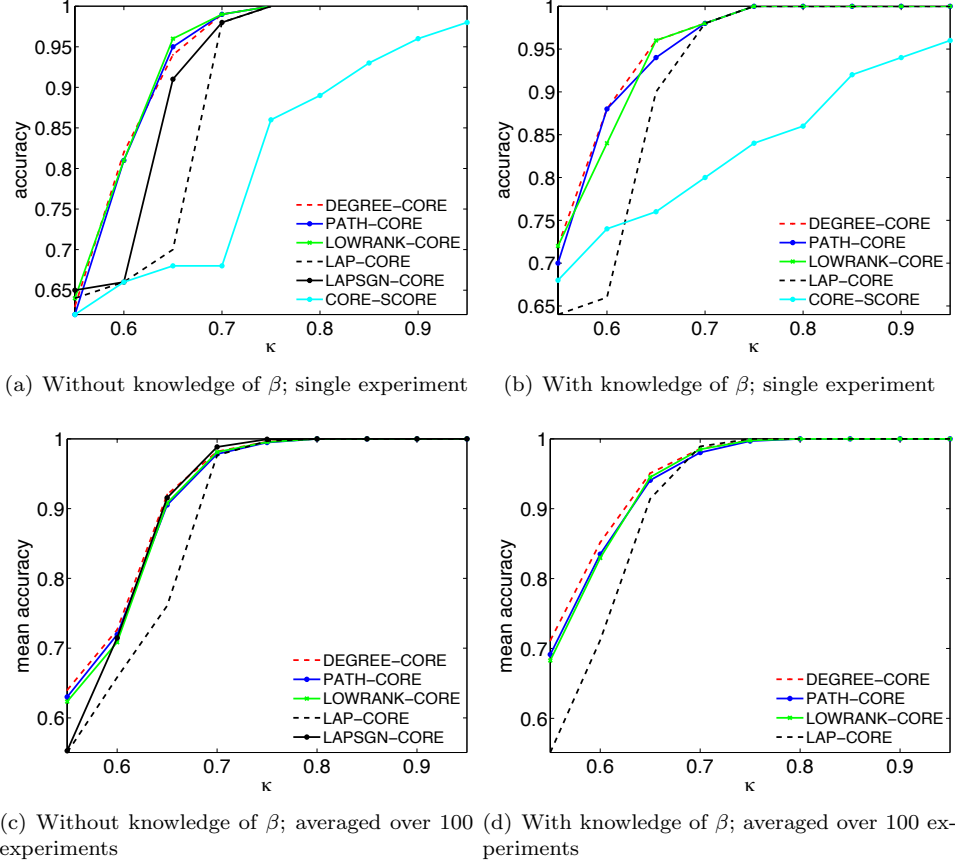


FIG. 10. Comparison of the methods for a graph with $n = 100$ vertices generated by a core-periphery block model with edge probabilities $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp}) = (\kappa, \kappa, 1 - \kappa)$ for $\kappa \in \{0.55, 0.60, \dots, 0.95\}$.

a vector of scores to measure coreness (using any of the above methods). Core-periphery structure is a common feature of real-world networks, and it is important to continue to develop methods to detect it and to compare the performance of such methods against each other on a wide variety of networks. We have introduced and explored the performance of several new methods in this paper. Importantly, the different methods that we introduce are based on rather different ideas, and it is very important to explore core-periphery structure from a multitude of perspectives.

Given the common use of k -cores in the consideration of core parts of networks, it is also interesting to examine the assignment of vertices into core and periphery sets based only on vertex degrees. Although using vertex degree as a measure of centrality or likelihood of belonging to a core can often produce inaccurate results [69], it can sometimes be true that a degree-based classification of vertices as core vertices or peripheral vertices should be successful for certain random-graph ensembles [85]. One can thus ask what properties such ensembles ought to have. More generally, it is also important to compare coreness scores with other centrality measures [5, 49, 69]. Another interesting question is whether one can use current methods for solving the group-synchronization problem (such as the eigenvector method and semidefinite programming [34, 37, 75]) for the detection of core-periphery structure in various families in networks.

An important future application is to examine core-periphery structure in temporal and multilayer networks [40, 46, 54]. Community structure (see, e.g., [41, 54]) has been studied in such contexts, and it should be very insightful to also consider core-periphery structure in multilayer networks. Another interesting direction is to develop additional objective functions with which to classify vertices into core and periphery sets.

Networks have many different types of mesoscale structures. In most research thus far, community structure has taken center stage. Other mesoscale structures, such as role assignment [71] and core-periphery structure [24], are also very important. These ideas are worthy of considerably more exploration.

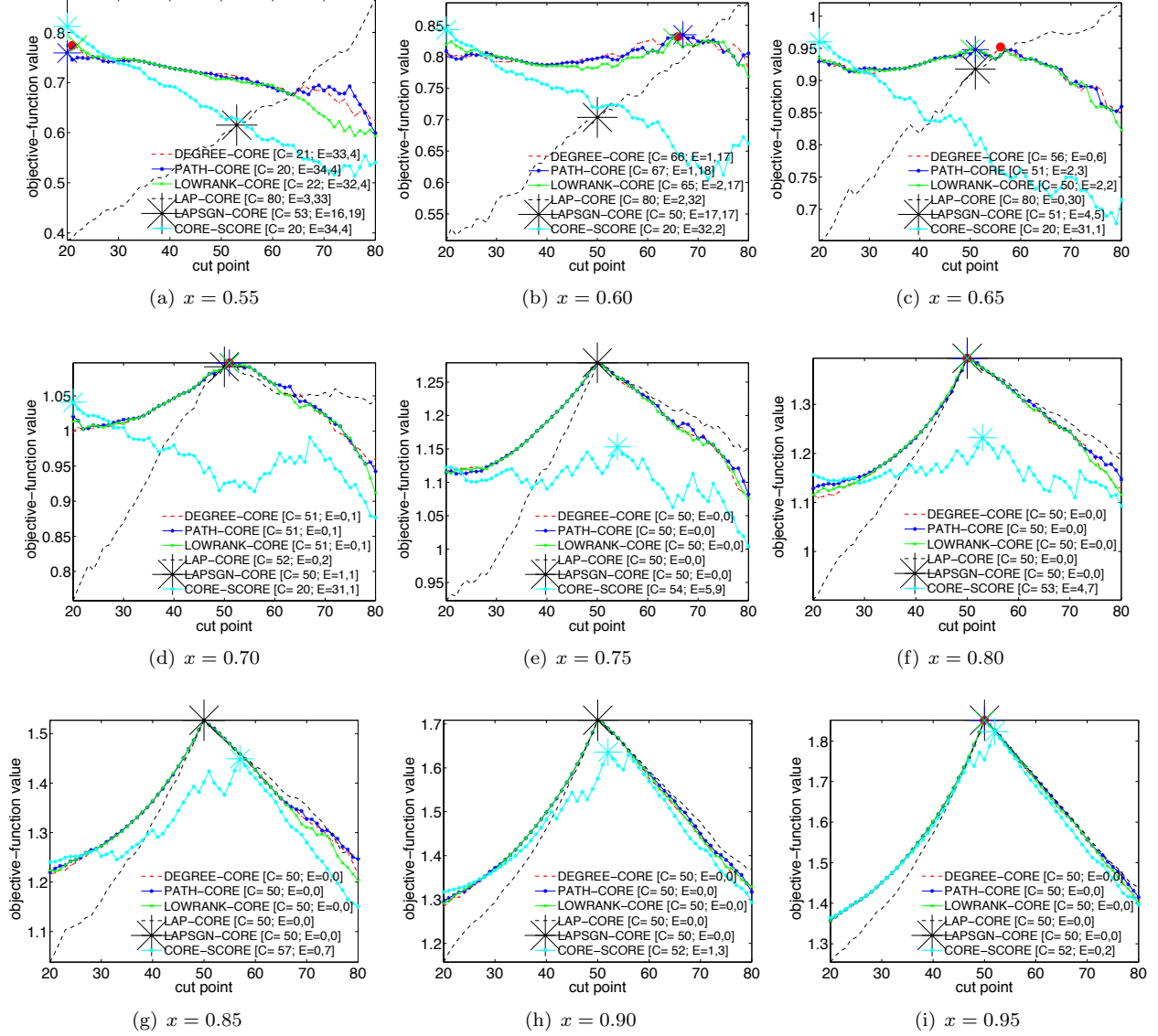


FIG. 11. Comparison of the values of the objective function (8) of the partition of networks into a core set and a periphery set. We calculate these values from the sorted scores from the various methods for detecting core-periphery structure as we vary the parameter κ in the ensemble $G(p_{cc}, p_{cp}, p_{pp})$ from Table 1 with $n = 100$. The probability vector in the block model is $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp}) = (\kappa, \kappa, 1 - \kappa)$. The “cut point” refers to the number of vertices in the core set. In the legends, C denotes the size of the core set that maximizes the objective function Eq. (8), and $\mathbf{E} = (y_1, y_2)$ denotes the corresponding 2-vector of errors. The first component of \mathbf{E} indicates the number of core vertices that we label as peripheral vertices, and the second indicates the number of peripheral vertices that we label as core vertices.

Acknowledgements. S.H.L. and M.A.P. were supported by a grant (EP/J001795/1) from the Engineering and Physical Sciences Research Council (EPSRC), and M.A.P. and P.R. were supported by the James S. McDonnell Foundation (#220020177). S.H.L. was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2011947). M.C. thanks Radek Erban and OCCAM at University of Oxford for their warm hospitality while hosting him for two months during Spring 2012 (during which this project was initiated) and is grateful to Amit Singer for his guidance and support via Award Number R01GM090200 from the NIGMS and Award Number FA9550-09-1-0551 from AFOSR. M.C. and P.R. also acknowledge support from AFOSR MURI grant FA9550-10-1-0569. P.R. acknowledges support from ONR grant N000141210040. Part of this work was undertaken while M.C. and P.R. were attending the Semester Program on Network Science and Graph Algorithms at the Institute for Computational and Experimental Research

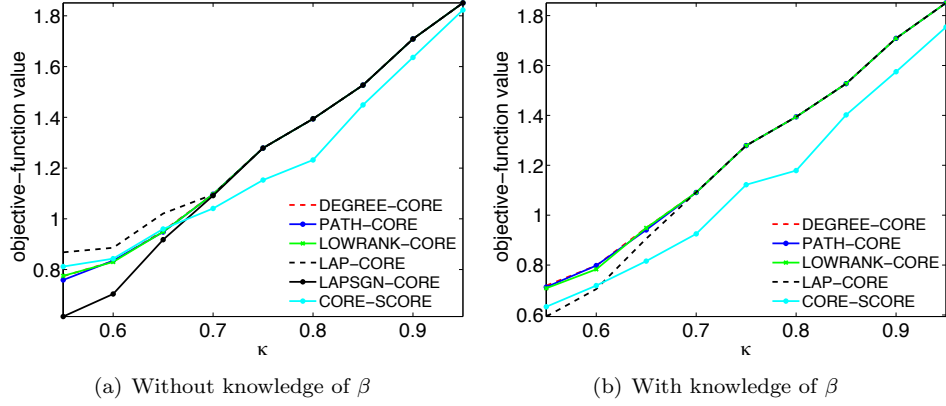


FIG. 12. Comparison of the actual values of the objective function (8) for a single experiment for all methods as a function of the parameter x in the ensemble $G(p_{cc}, p_{cp}, p_{pp})$ from Table 1 with $n = 100$ and $\beta = 0.5$. The probability vector in the block model is $\mathbf{p} = (p_{cc}, p_{cp}, p_{pp}) = (x, x, 1 - x)$.

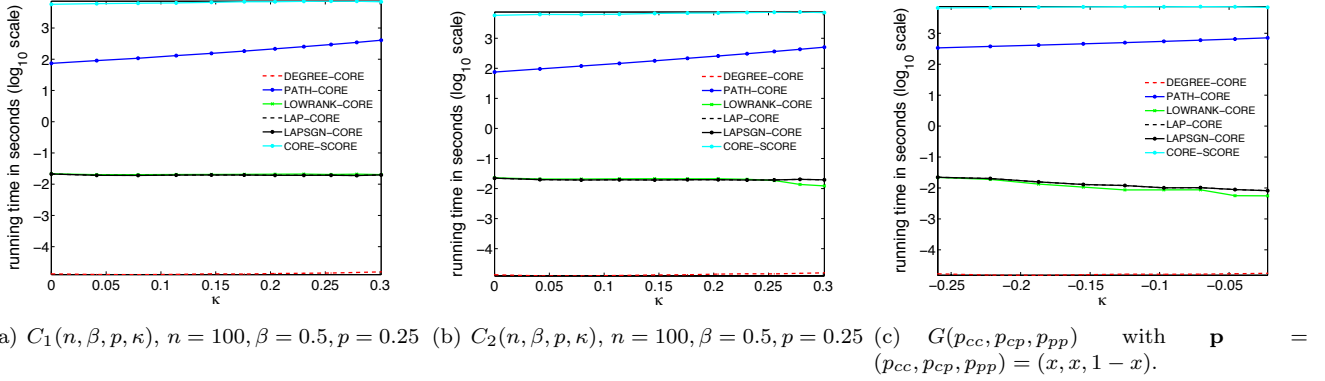


FIG. 13. Comparison of the computation times (in seconds and on a \log_{10} scale) for all methods and for three synthetic graph ensembles.

in Mathematics (ICERM) at Brown University.

REFERENCES

- [1] Y.-Y. AHN, J. P. BAGROW, AND S. LEHMANN, *Link communities reveal multiscale complexity in networks*, Nature, 466 (2010), pp. 761–764.
- [2] J. M. ANTHONISSE, *The Rush in a Directed Graph*, Stichting Mathematisch Centrum, Amsterdam, Oct. 1971. (notes, available at <http://oai.cwi.nl/oai/asset/9791/9791A.pdf>).
- [3] A. ARENAS, A. DÍAZ-GUILERA, AND C. J. PÉREZ-VICENTE, *Synchronization reveals topological scales in complex networks*, Phys. Rev. Lett., 96 (2006), p. 114102.
- [4] B. BALL, B. KARRER, AND M. E. J. NEWMAN, *Efficient and principled method for detecting communities in networks*, Phys. Rev. E, 84 (2011), p. 036103.
- [5] P. BARUCCA, D. TANTARI, AND F. LILLO, *Centrality metrics and localization in core-periphery networks*, arXiv:1510.01116v2.
- [6] J. BASCOMPTE, P. JORDANO, CARLOS J. MELIÁN, AND J. M. OLESEN, *The nested assembly of plant-animal mutualistic networks*, Proc. Natl. Acad. Sci. U.S.A., 100 (2003), pp. 9383–9387.
- [7] D. S. BASSETT, N. F. WYMBS, M. A. PORTER, P. J. MUCHA, J. M. CARLSON, AND S. T. GRAFTON, *Dynamic reconfiguration of human brain networks during learning*, Proc. Natl. Acad. Sci. U.S.A., 108 (2011), pp. 7641–7646.
- [8] D. S. BASSETT, N. F. WYMBS, M. P. ROMBACH, M. A. PORTER, P. J. MUCHA, AND S. T. GRAFTON, *Task-based core-periphery organization of human brain dynamics*, PLoS Comput. Biol., 9 (2013), p. e1003171.
- [9] M. BELKIN AND P. NIYOGI, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Comput., 15 (2003), pp. 1373–1396.
- [10] F. BENAYCH-GEORGES AND R. R. NADAKUDITI, *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Adv. Math., 227 (2011), pp. 494–521.
- [11] R. BHATIA, *Matrix Analysis*, vol. 169 of Graduate Texts in Mathematics, Springer Verlag, 1997.

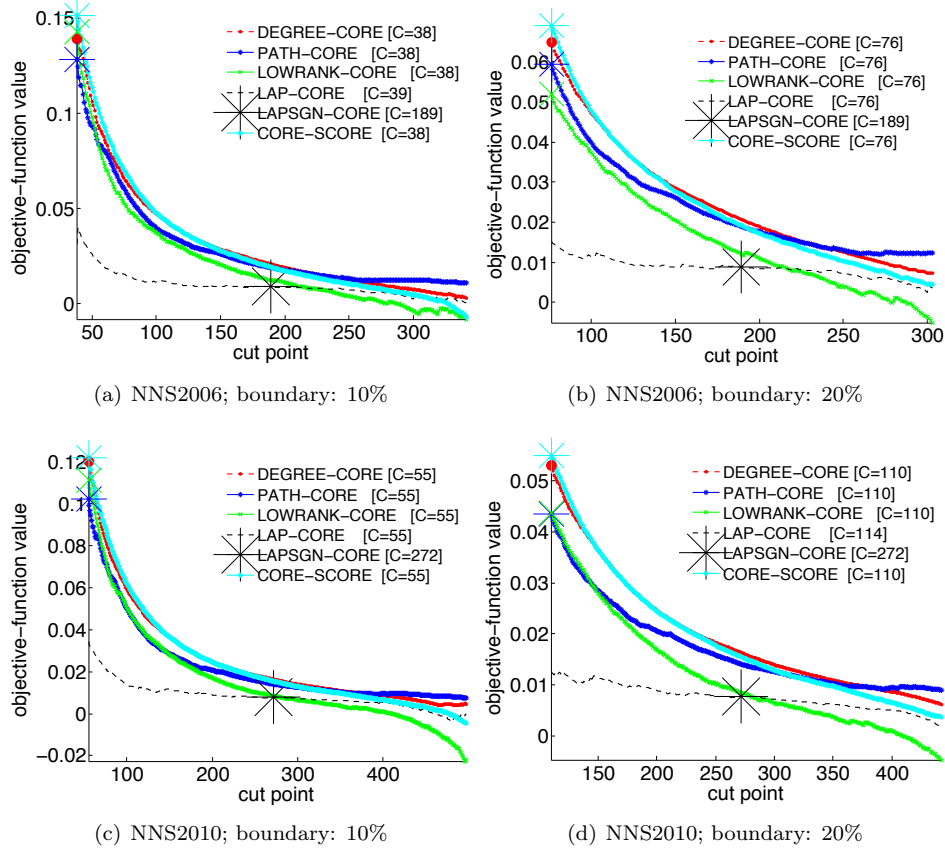


FIG. 14. Comparison of the methods for detecting core-periphery structure for networks of network scientists in (a,b) 2006 [57] and (c,d) 2010 [28] for the objective function in Eq. (8). We assume a minimum size for the core and periphery sets of at least (a,c) 10% of the vertices and (b,d) 20% of the vertices. We mark the cut points that maximize the objective functions on the curves as a large asterisk for LAPSGN-CORE and using other symbols for the other methods. The cut point refers to the number of core vertices, and the C values in the legends are the cut points that maximize the objective function (8). In other words, the optimal solution places C vertices in the core set.

- [12] S. P. BORGATTI AND M. G. EVERETT, *Models of core/periphery structures*, Soc. Networks, 21 (1999), pp. 375–395.
- [13] S. P. BORGATTI, M. G. EVERETT, AND L. C. FREEMAN, *UCInet*. Version 6.289, available at <http://www.analytictech.com/ucinet/>, 2011.
- [14] J. P. BOYD, W. J. FITZGERALD, M. C. MAHUTGA, AND D. A. SMITH, *Computing continuous core/periphery structures for social relations data with MINRES/SVD*, Soc. Networks, 32 (2010), pp. 125–137.
- [15] U. BRANDES, *A faster algorithm for betweenness centrality*, J. Math. Sociol., 25 (2001), pp. 163–177.
- [16] U. BRANDES, D. DELLING, M. GAERTLER, R. GÖRKE, M. HOEFER, Z. NIKOLOSKI, AND D. WAGNER, *On modularity clustering*, IEEE Trans. Knowl. Data Eng., 20 (2008), pp. 172–188.
- [17] C. CHASE-DUNN, *Global Formation: Structures of the World-Economy*, Basil Blackwell, Oxford, UK, 1989.
- [18] J. CHEN AND B. YUAN, *Detecting functional modules in the yeast protein-protein interaction network*, Bioinformatics, 22 (2006), pp. 2283–2290.
- [19] F. R. K. CHUNG, *Spectral Graph Theory*, Am. Math. Soc., 1997.
- [20] A. CLAUSET, S. ARBESMAN, AND D. B. LARREMORE, *Systematic inequality and hierarchy in faculty hiring networks*, Science Advances, 1 (2015).
- [21] R. R. COIFMAN AND S. LAFON, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.
- [22] R. R. COIFMAN, S. LAFON, A. B. LEE, M. MAGGIONI, B. NADLER, F. WARNER, AND S. W. ZUCKER, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, Proc. Natl. Acad. Sci. U.S.A., 102 (2005), pp. 7426–7431.
- [23] A. L. COMREY, *The minimum residual method of factor analysis*, Psychol. Rep., 11 (1962), pp. 15–18.
- [24] P. CSERMELY, A. LONDON, L.-Y. WU, AND B. UZZI, *Structure and dynamics of core/periphery networks*, J. Cplx. Netw., 1 (2013), pp. 93–123.
- [25] M. R. DA SILVA, H. MA, AND A.-P. ZENG, *Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks*, Proc. IEEE, 96 (2008), pp. 1411–1420.
- [26] R. B. DARLINGTON, S. L. WEINBERG, AND H. J. WALBERG, *Canonical variate analysis and related techniques*, Rev. Educ. Res., 43 (1973), pp. 433–454.
- [27] P. DOREIAN, *Structural equivalence in a psychology journal network*, J. Assoc. Inf. Sci., 36 (1985), pp. 411–417.

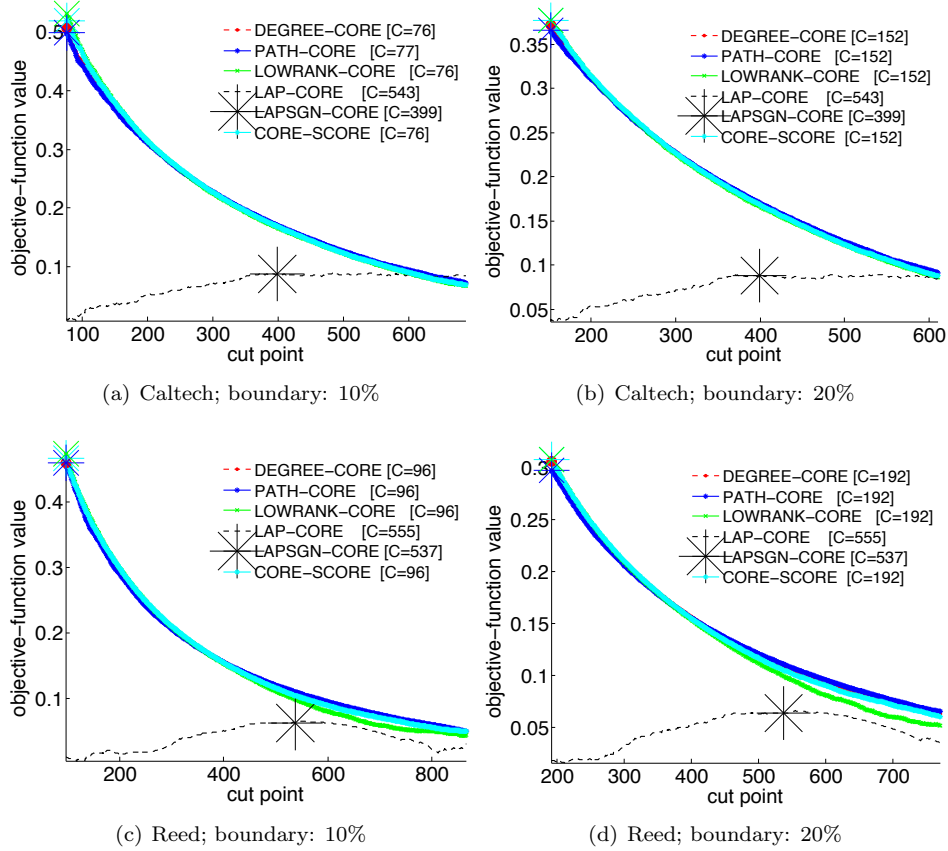


FIG. 15. Comparison of the methods for detecting core-periphery structure for Facebook networks [80, 81] of (a,b) Caltech and (c,d) Reed College for the objective function in Eq. (8). We assume a minimum size for the core and periphery sets of at least (a,c) 10% of vertices and (b,d) 20% of vertices. We mark the cut points that maximize the objective functions on the curves as a large asterisk for LAPSGN-CORE and using other symbols for the other methods. The cut point refers to the number of core vertices, and the C values in the legends are the cut point that maximizes the objective function (8). In other words, the optimal solution places C vertices in the core set.

- [28] D. EDLER AND M. ROSVALL, *The map generator software package (2010 network scientist coauthorship network)*. <http://mapequation.org/downloads/netscicoauthor2010.net>, 2010. [Online; accessed 12-September-2014].
- [29] M. G. EVERETT AND T. W. VALENTE, *Bridging, brokerage and betweenness*, Soc. Netw., 44 (2016), pp. 202–208.
- [30] D. FÉRAL AND S. PÉCHÉ, *The largest eigenvalue of rank one deformation of large Wigner matrices*, Comm. Math. Phys., 272 (2007), pp. 185–228.
- [31] S. FORTUNATO, *Community detection in graphs*, Phys. Rep., 486 (2010), pp. 75–174.
- [32] L. C. FREEMAN, *A set of measures of centrality based on betweenness*, 1977.
- [33] M. GIRVAN AND M. E. J. NEWMAN, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. USA, 99 (2002), pp. 7821–7826.
- [34] M. X. GOEMANS AND D. P. WILLIAMSON, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM, 42 (1995), pp. 1115–1145.
- [35] M. C. GONZÁLEZ, H. J. HERRMANN, J. KERTÉSZ, AND T. VICSEK, *Community structure and ethnic preferences in school friendship networks*, Physica A, 379 (2007), pp. 307–316.
- [36] B. H. GOOD, Y.-A. DE MONTJOYE, AND A. CLAUSET, *Performance of modularity maximization in practical contexts*, Phys. Rev. E, 81 (2010), p. 046106.
- [37] M. GRANT AND S. BOYD, *Graph implementations for nonsmooth convex programs*, in Recent Advances in Learning and Control, V. Blondel, S. Boyd, and H. Kimura, eds., Lecture Notes Contr. Inf. Sci., Springer-Verlag Limited, 2008, pp. 95–110. http://stanford.edu/~boyd/graph_dcp.html.
- [38] S. GUATTERY AND G. L. MILLER, *On the performance of spectral graph partitioning methods*, in Proc. 6th Ann. ACM-SIAM Symp. on Discrete Algorithms, SODA '95, Philadelphia, PA, USA, 1995, SIAM, pp. 233–242.
- [39] P. HOLME, *Core-periphery organization of complex networks*, Phys. Rev. E, 72 (2005), p. 046111.
- [40] P. HOLME AND J. SARAMÁKI, *Temporal networks*, Phys. Rep., 519 (2012), pp. 97–125.
- [41] L. G. S. JEUB, P. BALACHANDRAN, M. A. PORTER, P. J. MUCHA, AND M. W. MAHONEY, *Think locally, act locally: Detection of small, medium-sized, and large communities in large networks*, Phys. Rev. E, 91 (2015), p. 012821.
- [42] A. K. KELMANS, *The number of trees of a graph I*, Aut. and Remote Contr., 26 (1965), pp. 2118–2129.

- [43] ———, *The number of trees of a graph II*, Aut. and Remote Contr., 27 (1966), pp. 233–241.
- [44] ———, *Transformations of a graph increasing its Laplacian polynomial and number of spanning trees*, Europ. J. Comb., 18 (1997), pp. 35–48.
- [45] M. KITSACK, L. K. GALLOS, S. HAVLIN, F. LILJEROS, L. MUCHNIK, H. E. STANLEY, AND H. A. MAKSE, *Identification of influential spreaders in complex networks*, Nat. Phys., 6 (2010), pp. 888–893.
- [46] M. KIVELÄ, A. ARENAS, M. BARTHELEMY, J. P. GLEESON, Y. MORENO, AND M. A. PORTER, *Multilayer networks*, J. Cplx. Netw., 2 (2014), pp. 203–271.
- [47] P. KRUGMAN, *The Self-Organizing Economy*, Oxford University Press, Oxford, UK, 1996.
- [48] E. O. LAUMANN AND F. U. PAPP, *Networks of Collective Action: A Perspective on Community Influence*, Academic Press, New York, NY, 1976.
- [49] S. H. LEE, M. CUCURINGU, AND M. A. PORTER, *Density-based and transport-based core-periphery structures in networks*, Phys. Rev. E, 89 (2014), p. 032810.
- [50] A. C. F. LEWIS, N. S. JONES, M. A. PORTER, AND C. M. DEANE, *The function of communities in protein interaction networks at multiple scales*, BMC Syst. Biol., 4 (2010), p. 100.
- [51] G. MCLACHLAN AND D. PEEL, *Finite Mixture Models*, Wiley-Interscience, Hoboken, NJ, 2000.
- [52] M. MEILÄ AND J. SHI, *A random walks view of spectral segmentation*, in AI and STATISTICS (AISTATS), 2001.
- [53] F. MORONE AND H. A. MAKSE, *Influence maximization in complex networks through optimal percolation*, Nature, 524 (2015), pp. 65–68.
- [54] P. J. MUCHA, T. RICHARDSON, K. MACON, M. A. PORTER, AND J.-P. ONNELA, *Community Structure in Time-Dependent, Multiscale, and Multiplex Networks*, Science, 328 (2010), pp. 876–878.
- [55] Y. NESTEROV, *Introductory Lectures on Convex Optimization: A Basic Course*, Applied optimization, Kluwer Academic Publ., Boston, Dordrecht, London, 2004.
- [56] M. E. J. NEWMAN, *A measure of betweenness centrality based on random walks*, Soc. Networks, 27 (2005), pp. 39–54.
- [57] ———, *Finding community structure in networks using the eigenvectors of matrices*, Phys. Rev. E, 74 (2006), p. 036104.
- [58] ———, *Networks: An Introduction*, Oxford University Press, USA, 2010.
- [59] M. E. J. NEWMAN AND M. GIRVAN, *Mixing patterns and community structure in networks*, in Statistical Mechanics of Complex Networks, Romualdo Pastor-Satorras, Miguel Rubi, and Albert Diaz-Guilera, eds., vol. 625 of Lecture Notes in Physics, Springer Berlin Heidelberg, 2003, pp. 66–87.
- [60] ———, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), p. 026113.
- [61] J.-P. ONNELA, J. SARAMÄKI, J. HYVÖNEN, G. SZABÓ, D. LAZER, K. KASKI, J. KERTÉSZ, AND A. L. BARABÁSI, *Structure and tie strengths in mobile communication networks*, Proc. Natl. Acad. Sci. U.S.A., 104 (2007), pp. 7332–7336.
- [62] G. PALLA, I. DERENYI, I. FARKAS, AND T. VICSEK, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, 435 (2005), pp. 814–818.
- [63] TIAGO P. PEIXOTO, *Hierarchical block structures and high-resolution model selection in large networks*, Phys. Rev. X, 4 (2014), p. 011047.
- [64] C. PICCARDI, *Finding and testing network communities by lumped Markov chains*, PLoS ONE, 6 (2011), p. e27028.
- [65] P. PONS AND M. LATAPY, *Computing communities in large networks using random walks*, in Comp. and Inf. Sci. - ISCIS 2005, vol. 3733 of Lecture Notes in Computer Science, Springer, 2005, pp. 284–293.
- [66] M. A. PORTER, P. J. MUCHA, M. E. J. NEWMAN, AND C. M. WARMBRAND, *A network analysis of committees in the U.S. House of Representatives*, Proc. Natl. Acad. Sci. U.S.A., 102 (2005), pp. 7057–7062.
- [67] M. A. PORTER, J.-P. ONNELA, AND P. J. MUCHA, *Communities in networks*, Notices Amer. Math. Soc., 56 (2009), pp. 1082–1097, 1164–1166.
- [68] T. RICHARDSON, P. J. MUCHA, AND M. A. PORTER, *Spectral tripartitioning of networks*, Phys. Rev. E, 80 (2009), p. 036111.
- [69] M. P. ROMBACH, M. A. PORTER, J. H. FOWLER, AND P. J. MUCHA, *Core-periphery structure in networks*, SIAM J. Appl. Math., 74 (2014), pp. 167–190.
- [70] F. D. ROSSA, F. DERCOLE, AND C. PICCARDI, *Profiling core-periphery network structure by random walkers*, Sci. Rep., 3 (2013).
- [71] R. A. ROSSI AND N. K. AHMED, *Role discovery in networks*. arXiv:1405.7134v2, 2014.
- [72] M. ROSVALL AND C. T. BERGSTROM, *Maps of random walks on complex networks reveal community structure*, Proc. Natl. Acad. Sci. U.S.A., 105 (2008), pp. 1118–1123.
- [73] S. T. ROWEIS AND L. K. SAUL, *Nonlinear dimensionality reduction by locally linear embedding*, Science, 290 (2000), pp. 2323–2326.
- [74] J. SHI AND J. MALIK, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888–905.
- [75] A. SINGER, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal., 30 (2011), pp. 20–36.
- [76] D. A. SMITH AND D. R. WHITE, *Structure and dynamics of the global economy: Network analysis of international trade*, Social Forces, 70 (1992), pp. 857–893.
- [77] D. A. SPIELMAN AND S.-H. TENG, *Spectral partitioning works: Planar graphs and finite element meshes.*, in FOCS, IEEE Computer Society, 1996, pp. 96–105.
- [78] ———, *Spectral partitioning works: Planar graphs and finite element meshes*, Linear Algebra Appl., 421 (2007), pp. 284–305. Special Issue in honor of Miroslav Fiedler.
- [79] S. STEIBER, *The world system and world trade: An empirical explanation of conceptual conflicts*, Sociol. Q., 20 (1979), pp. 23–26.
- [80] A. L. TRAUD, E. D. KELSIC, P. J. MUCHA, AND M. A. PORTER, *Comparing Community Structure to Characteristics in Online Collegiate Soc. Netw.*, SIAM Rev., 53 (2011), pp. 526–543.
- [81] A. L. TRAUD, P. J. MUCHA, AND M. A. PORTER, *Social structure of facebook networks*, Physica A, 391 (2012), pp. 4165–4180.
- [82] T. W. VALENTE AND K. FUJIMOTO, *Bridging: Locating critical connectors in a network*, Soc. Netw., 32 (2010), pp. 212–220.
- [83] I. WALLERSTEIN, *The Modern World-System*, Academic Press, New York, NY, 1974.
- [84] J. YANG AND J. LESKOVEC, *Structure and overlaps of ground-truth communities in networks*, ACM Trans. Intell. Syst. Technol., 5 (2014), pp. 26:1–26:35.
- [85] X. ZHANG, T. MARTIN, AND M. E. J. NEWMAN, *Identification of core-periphery structure in networks*, Phys. Rev. E, 91 (2015), p. 032803.

Appendix 1: Algorithms for Computing Path-Core scores. In this appendix, we detail the steps of our proposed algorithm for computing the PATH-CORE scores for all vertices in an unweighted graph G .

Algorithm 5 PATH-CORE: Computes PATH-CORE scores for all vertices of a graph G .

<p>Input: G Output: C_P</p> <p>1: $C_P(w) \leftarrow 0, w \in V$;</p> <p>2: for $(s, t) \in E(G)$ do</p> <p>3: $G' \leftarrow G \setminus (s, t)$;</p> <p>4: $\sigma_s(w), \sigma_t(w) \leftarrow 0, v \in V; \sigma_s(s), \sigma_t(t) \leftarrow 1$;</p> <p>5: $d_s(w), d_t(w) \leftarrow -1, v \in V; d_s(s), d_t(t) \leftarrow 0$;</p> <p>6: $Q \leftarrow$ empty queue;</p> <p>7: enqueue $s \rightarrow Q$;</p> <p>8: while Q not empty do</p> <p>9: dequeue $w \leftarrow Q$;</p> <p>10: for each $u \in \Gamma_{G'}(w)$ do</p> <p>11: $(d_s, Q) \leftarrow \text{DISTANCE}(w, u, d_s, Q)$</p> <p>12: end for</p> <p>13: end while</p> <p>14: enqueue $t \rightarrow Q$;</p> <p>15: while Q not empty do</p> <p>16: dequeue $w \leftarrow Q$;</p>	<p>17: for each $u \in \Gamma_{G'}(w)$ do</p> <p>18: $(d_t, Q) \leftarrow \text{DISTANCE}(w, u, d_t, Q)$</p> <p>19: $\sigma_t \leftarrow \text{PATH-COUNT}(w, u, d_s, \sigma_t)$</p> <p>20: end for</p> <p>21: end while</p> <p>22: enqueue $s \rightarrow Q$;</p> <p>23: while Q not empty do</p> <p>24: dequeue $w \leftarrow Q$;</p> <p>25: for each $u \in \Gamma_{G'}(w)$ do</p> <p>26: if $d_t(w) < d_t(u)$ then</p> <p>27: enqueue $u \rightarrow Q$;</p> <p>28: end if</p> <p>29: $\sigma_s \leftarrow \text{PATH-COUNT}(w, u, d_t, \sigma_s)$</p> <p>30: end for</p> <p>31: end while</p> <p>32: for $w \in V \setminus (s, t)$ do</p> <p>33: $C_P(w) = C_P(w) + \sigma_s(w) \cdot \sigma_t(w) / \sigma_s(t)$;</p> <p>34: end for</p> <p>35: end for</p>
--	---

LEMMA 8.1. *Algorithm 5 outputs the PATH-CORE scores for all vertices in an unweighted graph G .*

Proof. It suffices to show for one edge $(s, t) \in E(G)$ and one iteration (i.e., lines 3–33) that the algorithm counts, for each vertex $w \in V(G) \setminus (s, t)$, the number of shortest paths between s and t that contain w . This number $\sigma_{s,t}(w)$ is given by the algorithm as $\sigma_s(w) \cdot \sigma_t(w)$. In this case, $\sigma_s(w)$ is the number of paths between s and w of length $d_{G'}(s, w)$ [where the graph $G' = G \setminus (s, t)$] if and only if w lies on a shortest path between s and t .

Algorithm 5 performs three breadth-first-searches (BFSs). In the first BFS, it searches from vertex s and records the distances from s to all other vertices. It then performs a BFS starting from vertex t . During this second BFS, it records the distances to all vertices from t , and it also records $\sigma_t(w)$ for vertices that lie on a shortest path between s and t . The PATH-SCORE algorithm knows that u lies on a shortest path between s and t if it has a distance from s that is less than the distance from s of its predecessor in the BFS from t . In other words, if $d_t(w) < d_t(u)$, then an edge (w, u) lies on a shortest path between s and t if and only if $d_s(u) < d_s(w)$. Additionally,

$$\sigma_t(u) = \sum_{w \in P_t(u)} \sigma_t(w).$$

In the second BFS, Algorithm 5 finds a vertex u exactly once for each of its predecessors $w \in P_t(u)$, and it adds $\sigma_t(w)$ to $\sigma_t(u)$. Therefore, in the second BFS, for each vertex $u \in V(G) \setminus (s, t)$, PATH-SCORE records $\sigma_t(u)$ as the number of shortest paths from t to u if u is on a shortest path between s and t . If it is not, then $\sigma_t(u)$ is still 0.

By the same arguments, in the third BFS, for each vertex $u \in V(G) \setminus (s, t)$, PATH-CORE records $\sigma_s(u)$ as the number of shortest paths from s to u if u is on a shortest path between s and t . If it is not, then $\sigma_s(u)$ is still 0.

It should now be clear that, for all $w \in V(G) \setminus (s, t)$, it follows that $\sigma_s(w) \cdot \sigma_t(w)$ yields $\sigma_{s,t}(w)$. \square

LEMMA 8.2. *Algorithm 5 finishes in $\mathcal{O}(m^2)$ time.*

Proof. Algorithm 5 iterates (i.e., it runs lines 3–33) once for each edge. In one iteration, it performs three BFSs. During a BFS, every edge of G' is considered exactly once; for each edge, it employs the subroutines DISTANCE, PATH-COUNT, or both. Because DISTANCE and PATH-COUNT both run in $\mathcal{O}(1)$ time, every iteration of PATH-CORE runs in $\mathcal{O}(m)$ time. Therefore, the temporal complexity of PATH-CORE is $\mathcal{O}(m^2)$. \square

For weighted graphs, one can implement an algorithm that is very similar to Algorithm 5. This algorithm uses Dijkstra's algorithm for shortest paths instead of BFS, and it runs in $\mathcal{O}(m + n \log n)$ time instead of $\mathcal{O}(m)$, so the total temporal complexity becomes $\mathcal{O}(m^2 + mn \log n)$.

Algorithm 6 DISTANCE: Records distance from a given vertex for unvisited vertices in the breadth-first search.

Input: w, u, d, Q
Output: d, Q
1: **if** $d(u) < 0$ **then**
2: enqueue $u \rightarrow Q$;
3: $d(w) \leftarrow d(w) + 1$;
4: **end if**

Algorithm 7 PATH-COUNT: Records the number of shortest paths through u .

Input: w, u, d, σ
Output: σ
1: **if** $d(u) < d(w)$ **then**
2: $\sigma(u) = \sigma(u) + \sigma(w)$;
3: **end if**

Appendix 2: Symmetry in the Random-Walk Laplacian. We now show that a symmetry relation like (17) exists for the random-walk Laplacian only under certain conditions. Additionally, the most obvious version of such a statement does not hold. To see this, let \mathbf{x} be an eigenvector of \bar{L} (which is nontrivial, so $\mathbf{x} \perp \mathbf{1}_n$). We use the notation $\bar{D} = \text{diag}(n - 1 - d_i)$, where d_i denotes the degree of vertex i , and calculate

$$\begin{aligned}
\bar{L} &= \bar{D}^{-1} \bar{G} \\
&= \bar{D}^{-1} (J_n - G - I_n) \\
&= \bar{D}^{-1} (J_n - I_n) - \bar{D}^{-1} G \\
&= \bar{D}^{-1} (J_n - I_n) - \bar{D}^{-1} D D^{-1} G \\
&= \bar{D}^{-1} (J_n - I_n) - \bar{D}^{-1} D L.
\end{aligned}
\tag{19}$$

Because $\bar{D}^{-1} = \text{diag}\left(\frac{1}{n-1-d_i}\right)$ and $D = \text{diag}(d_i)$ are diagonal matrices, it follows that $\bar{D}^{-1} D = \text{diag}\left(\frac{d_i}{n-1-d_i}\right)$. Given an eigenvector \mathbf{x} of \bar{L} , it follows that $\bar{L}\mathbf{x} = \bar{\lambda}\mathbf{x}$ for some eigenvalue $\bar{\lambda}$. Because \bar{L} is a row-stochastic matrix, it has the trivial eigenvalue $\bar{\lambda}_1 = 1$ with associated eigenvector $\bar{\mathbf{v}}_1 = \mathbf{1}_n$. We apply both sides of Eq. (19) to the eigenvector \mathbf{x} and note that $J_n \mathbf{x} = \mathbf{0}$ because $\mathbf{x} \perp \bar{\mathbf{v}}_1 = \mathbf{1}_n$. We thereby obtain

$$\begin{aligned}
\bar{\lambda}\mathbf{x} &= \bar{L}\mathbf{x} \\
&= \bar{D}^{-1} (J_n - I_n) \mathbf{x} - \bar{D}^{-1} D L \mathbf{x} \\
&= -\bar{D}^{-1} I_n \mathbf{x} - \bar{D}^{-1} D L \mathbf{x}.
\end{aligned}
\tag{20}$$

Multiplying both sides of Eq. (20) by $D^{-1} \bar{D}$ on the left yields

$$D^{-1} \bar{D} \bar{\lambda} \mathbf{x} = -D^{-1} \bar{D} \bar{D}^{-1} \mathbf{x} - L \mathbf{x}, \tag{21}$$

so

$$\begin{aligned}
L \mathbf{x} &= -D^{-1} \bar{D} \bar{\lambda} \mathbf{x} - D^{-1} \mathbf{x} \\
&= -(\bar{D} \bar{\lambda} + I_n) D^{-1} \mathbf{x} \\
&= -\text{diag}\left(\frac{\bar{\lambda}(n-1-d_i) + 1}{d_i}\right) \mathbf{x}.
\end{aligned}$$

Therefore, \mathbf{x} is not an eigenvector of L unless

$$\frac{\bar{\lambda}(n-1-d_i) + 1}{d_i} = \theta \quad \text{for all } i \in \{1, \dots, n\},$$

for some constant θ . In other words, $d_i = \frac{\bar{\lambda}(n-1)}{\theta + \bar{\lambda}}$ for $i \in \{1, \dots, n\}$, so the graph is d -regular, with $d_1 = d_2 = \dots = d_n = d = \frac{\bar{\lambda}(n-1)}{\theta + \bar{\lambda}}$. Therefore,

$$\theta = \frac{\bar{\lambda}(n-1-d) + 1}{d},$$

so the eigenvector \mathbf{x} of \bar{L} is also an eigenvector of L (with a corresponding eigenvalue of $-\theta$).

Appendix 3: Planted High-Degree Vertices. To illustrate the sensitivity of the DEGREE-CORE method to the presence of high-degree peripheral vertices, we conduct a numerical experiment in which we intentionally plant high-degree vertices in the periphery set. This helps illustrate that it is dangerous to use methods like k -core decomposition (which has very strong demands that vertices have a high degree to be construed as core vertices) to study core-periphery structure [24]. In Fig. 16, we consider a graph from the ensemble $G(p_{cc}, p_{cp}, p_{pp}, n_c, n_p)$ with $n = 100$ vertices, edge probabilities $(p_{cc}, p_{cp}, p_{pp}) = (0.4, 0.4, 0.2)$, n_c core vertices, n_p peripheral vertices (with $n = n_c + n_p$), and planted high-degree vertices in the periphery set. To perturb the graph G from the above ensemble to plant high-degree peripheral vertices, we proceed as follows. First, we select each peripheral vertex with independent probability 0.1. Second, we connect each such vertex to 15 non-neighboring peripheral vertices that we choose uniformly at random. In the left panel of Fig. 16, we show an example with a boundary size of 10% boundary size, so we are assuming that the core and periphery sets each have at least $0.1n = 10$ vertices. We search for a cut point in the interval $[10, 90]$. In the right panel, we consider a larger boundary size and assume that the core and the periphery sets each have at least 25 vertices. We now search for an optimal cut in the interval $[25, 75]$. In the two planted-degree scenarios for which the size of the core set is unknown, all methods yield many misclassified vertices, although the LAPSGN-CORE method has the lowest number (28) of misclassifications in both cases.

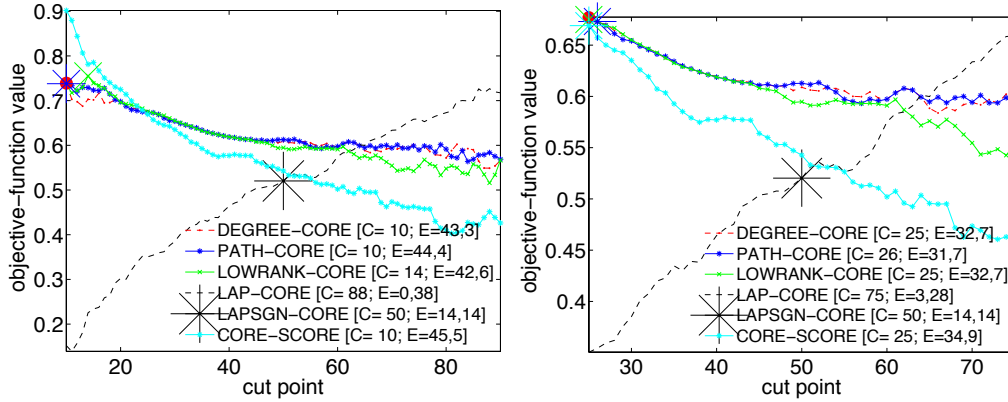


FIG. 16. Comparison of the methods for one realization of the graph ensemble $G(p_{cc}, p_{cp}, p_{pp}, n, n_c, n_p)$ with $n = 100$ vertices, $n_c = 50$ core vertices, $n_p = 50$ peripheral vertices, edge probabilities $(p_{cc}, p_{cp}, p_{pp}) = (0.4, 0.4, 0.2)$, and planted high-degree vertices for the objective function in Eq. (8). The cut point refers to the number of core vertices. In the legends, C denotes the size of the core set that maximizes the objective function Eq. (8), and $\mathbf{E} = (y_1, y_2)$ denotes the corresponding 2-vector of errors. The first component of \mathbf{E} indicates the number of core vertices that we label as peripheral vertices, and the second indicates the number of peripheral vertices that we label as core vertices. In this graph, each peripheral vertex has a probability of 0.1 of becoming adjacent to 15 additional non-neighboring peripheral vertices that we select uniformly at random. We mark the cut points that maximize the objective functions on the curves as a large asterisk for LAPSGN-CORE and using other symbols whose colors match the colors of the corresponding curves for the other methods. The cut point refers to the number of core vertices.

η	DEGREE-CORE	PATH-CORE	LOWRANK-CORE	LAP-CORE	LAPSGN-CORE	CORE-SCORE
$ V_C \geq 10, V_P \geq 10$	43+3	44+4	42+6	0+38	14+14	45+5
$ V_C \geq 25, V_P \geq 25$	32+7	31+7	32+7	3+28	14+14	34+9
Known $\beta = 0.5$	11+11	11+11	12+12	14+14	14+14	19+19

TABLE 3

Total number of misclassifications by each method for a graph from the ensemble $G(p_{cc}, p_{cp}, p_{pp}, n)$ with $n = 200$ vertices, edge probabilities $(p_{cc}, p_{cp}, p_{pp}) = (0.4, 0.4, 0.2)$, and planted high-degree vertices. In this graph, each peripheral vertex has a probability of 0.1 of becoming adjacent to 15 additional non-neighboring peripheral vertices that we select uniformly at random. Each entry in the table records two types of errors: the first number indicates the number of core vertices misclassified as peripheral vertices, and the second number indicates the number of peripheral vertices that are misclassified as core vertices.

Appendix 4: Correlations Between Coreness Values from Different Methods. In Table 4, we consider several empirical networks and examine the numerical values of the Pearson and Spearman correlations between the coreness values that we obtain for the core-periphery detection methods that we examine.

NNS2006	q, \mathcal{C}	q, \mathcal{P}	q, \mathcal{R}	q, \mathcal{L}	\mathcal{C}, \mathcal{P}	\mathcal{C}, \mathcal{R}	\mathcal{C}, \mathcal{L}	\mathcal{P}, \mathcal{R}	\mathcal{P}, \mathcal{L}	\mathcal{R}, \mathcal{L}
Pearson	0.79*	0.89*	0.72*	0.02	0.64*	0.56*	0.03	0.62*	0.03	-0.01
Spearman	0.79*	0.62*	0.43*	0.04	0.37*	0.65*	-0.05	0.14*	-0.01	0.01
S_{frac}	0.93	0.90	0.79	0.69	0.88	0.82	0.69	0.78	0.69	0.67
(z-score)	(19.5 [†])	(17.4 [†])	(8.2 [†])	(1.1)	(15.7 [†])	(10.2 [†])	(1.1)	(7.5 [†])	(0.7)	(-0.9)
	q, \mathcal{LS}	$\mathcal{C}, \mathcal{LS}$	$\mathcal{P}, \mathcal{LS}$	$\mathcal{R}, \mathcal{LS}$	$\mathcal{L}, \mathcal{LS}$					
S_{frac}	0.50	0.47	0.47	0.44	0.70					
(z-score)	(0.1)	(-1.5)	(-1.9)	(-3.4 [†])	(11.9 [†])					
NNS2010	q, \mathcal{C}	q, \mathcal{P}	q, \mathcal{R}	q, \mathcal{L}	\mathcal{C}, \mathcal{P}	\mathcal{C}, \mathcal{R}	\mathcal{C}, \mathcal{L}	\mathcal{P}, \mathcal{R}	\mathcal{P}, \mathcal{L}	\mathcal{R}, \mathcal{L}
Pearson	0.78*	0.84*	0.71*	0.01	0.62*	0.46*	0.01	0.56*	0.02	0.01
Spearman	0.84*	0.56*	0.39*	0.10	0.38*	0.56*	0.04	0.17*	0.08	0.03
S_{frac}	0.96	0.88	0.80	0.71	0.87	0.82	0.72	0.76	0.70	0.75
(z-score)	(29.1 [†])	(20.2 [†])	(12.4 [†])	(3.7 [†])	(19.8 [†])	(14.0 [†])	(4.1 [†])	(8.7 [†])	(1.9 [†])	(7.8 [†])
	q, \mathcal{LS}	$\mathcal{C}, \mathcal{LS}$	$\mathcal{P}, \mathcal{LS}$	$\mathcal{R}, \mathcal{LS}$	$\mathcal{L}, \mathcal{LS}$					
S_{frac}	0.54	0.52	0.50	0.51	0.71					
(z-score)	(2.7 [†])	(1.5)	(-0.1)	(0.9)	(16.9 [†])					
Caltech	q, \mathcal{C}	q, \mathcal{P}	q, \mathcal{R}	q, \mathcal{L}	\mathcal{C}, \mathcal{P}	\mathcal{C}, \mathcal{R}	\mathcal{C}, \mathcal{L}	\mathcal{P}, \mathcal{R}	\mathcal{P}, \mathcal{L}	\mathcal{R}, \mathcal{L}
Pearson	0.96*	0.97*	0.98*	0.02	0.86*	0.97*	0.01	0.93*	0.01	0.02
Spearman	1.00*	0.99*	0.99*	0.09	0.98*	1.00*	0.08	0.97*	0.09*	0.07
S_{frac}	0.98	0.97	0.97	0.43	0.97	0.99	0.42	0.96	0.42	0.42
(z-score)	(38.4 [†])	(36.9 [†])	(36.7 [†])	(5.7 [†])	(35.5 [†])	(38.4 [†])	(5.0 [†])	(34.4 [†])	(5.5 [†])	(4.7 [†])
	q, \mathcal{LS}	$\mathcal{C}, \mathcal{LS}$	$\mathcal{P}, \mathcal{LS}$	$\mathcal{R}, \mathcal{LS}$	$\mathcal{L}, \mathcal{LS}$					
S_{frac}	0.53	0.53	0.53	0.53	0.81					
(z-score)	(4.7 [†])	(4.2 [†])	(4.4 [†])	(4.1 [†])	(25.7 [†])					
Reed	q, \mathcal{C}	q, \mathcal{P}	q, \mathcal{R}	q, \mathcal{L}	\mathcal{C}, \mathcal{P}	\mathcal{C}, \mathcal{R}	\mathcal{C}, \mathcal{L}	\mathcal{P}, \mathcal{R}	\mathcal{P}, \mathcal{L}	\mathcal{R}, \mathcal{L}
Pearson	0.92*	0.95*	0.98*	-0.01	0.77*	0.94*	-0.02	0.90*	-0.01	-0.01
Spearman	0.99*	0.98*	0.96*	0.07	0.96*	0.98*	0.07	0.90*	0.09*	0.05
S_{frac}	0.99	0.98	0.98	0.51	0.97	0.99	0.51	0.96	0.51	0.51
(z-score)	(41.8 [†])	(39.7 [†])	(40.2 [†])	(6.3 [†])	(39.1 [†])	(42.0 [†])	(6.5 [†])	(38.2 [†])	(6.5 [†])	(6.3 [†])
	q, \mathcal{LS}	$\mathcal{C}, \mathcal{LS}$	$\mathcal{P}, \mathcal{LS}$	$\mathcal{R}, \mathcal{LS}$	$\mathcal{L}, \mathcal{LS}$					
S_{frac}	0.51	0.52	0.52	0.52	0.98					
(z-score)	(5.4 [†])	(5.9 [†])	(5.6 [†])	(5.7 [†])	(42.5 [†])					

TABLE 4

Pearson and Spearman correlation coefficients for various coreness measures and the similarity fraction S_{frac} for core-periphery partitioning with a boundary of 20% of the vertices (see the right panels in Figs. 14 and 15) between the objective function in Eq. (8) for several empirical networks. We use the notation q for DEGREE-CORE, \mathcal{C} for CORE-SCORE, \mathcal{P} for PATH-CORE, \mathcal{R} for LOWRANK-CORE, \mathcal{L} for LAP-CORE, and \mathcal{LS} for LAPSGN-CORE (for which we calculate only S_{frac} given by Eq. (18) because its score is the same as \mathcal{L}). We use the designation * for the correlation values that have a p -value smaller than 0.01 and the designation [†] for z -scores whose absolute value is larger than 2. We construe these results as statistically significant. We calculate the z -scores by randomly permuting the vertex indices (with 10000 different applications of such a permutation for each calculation) as described in [80]: $z = (S_{\text{frac}} - \mu)/(\text{stdev})$ where μ and “stdev” are the means and standard deviations of the S_{frac} values for random permutations.